Trademark Office ("the Office"). Furthermore, as these database citations concern a human genomic clone from chromosome 2 (GenBank Accession Number AC025750), and a cDNA clone from *Rattus norvegicus* (GenBank Accession Number AI043703), Applicants submit that these citations are not germane to the presently pending claims, and are therefore are not provided.

Second, the Examiner notes that "the Applicant has not submitted an English language translation or an English language abstract for document DE 19841413C" (the Action at page 2). Applicants note for the record that only the German patent application itself was provided by the European Patent Office, which is why no English language abstract was provided to the Office. Therefore, Applicants provide herewith a copy of the English language abstract from the published PCT patent application based on DE 19841413C (**Exhibit A**; see priority information).

IV. Title

The Action objects to the title of the application as allegedly "not descriptive" (the Action at page 3). Applicants have amended the title of the present application based on the suggestion from the Examiner.

Applicants request that, since the objection has been overcome, this objection be withdrawn.

V. Rejection of Claims 1, 3, and 5-11 Under 35 U.S.C. § 101

The Action first rejects claims 1, 3, and 5-11 under 35 U.S.C. § 101, as allegedly lacking a patentable utility. Applicants respectfully traverse.

First, while Applicants in no way agree with the Examiner's position that claims 1, 7 and 10 lack a patentable utility, as claims 1, 7 and 10 have been cancelled entirely without prejudice and without disclaimer, the present rejection of claims 1, 7 and 10 under 35 U.S.C. § 101 is rendered moot. The remainder of this section will therefore focus on claims 3, 5, 6, 8, 9 and 11.

The presently claimed sequence has clearly been described by Applicants in the specification as originally filed as an ion channel protein (see, at least, the title of the application as originally filed, and page 2, lines 2-4 of the specification), and more particularly a voltage-gated potassium channel protein (see, at least, page 2, line 5 of the specification). Additionally, Applicants respectfully point out that the presently claimed sequence shares **100% identity** at the amino acid level over the entire length of

SEQ ID NO:2 with two sequences that are present in the leading scientific repository for biological sequence data (GenBank), which have been annotated by independent third party scientists *wholly unaffiliated with Applicants* as "Homo sapiens voltage-gated potassium channel subunit Kv10.1a" (GenBank accession number AF454547; alignment and GenBank report provided in **Exhibit B**), and "Homo sapiens potassium voltage-gated channel subfamily G, member 3" (Kv6.3, GenBank accession number NM_172344; alignment and GenBank report provided in **Exhibit C**). It is well known in the art that the Kv10.1a and Kv6.3 subunits are alternative names for the same protein (see page 2 from the GenBank report provided in **Exhibit C**). Furthermore, three independent groups of scientists have established that the presently claimed sequence specifically interacts with the well-studied Kv2.1 voltage-gated potassium channel subunit to form functional voltage-gated potassium ion channels (Sano *et al.*, *FEBS Lett.* **512**:230-234, 2002 ("Sano"; copy of the abstract provided in **Exhibit D**); Ottschytsch *et al.*, *Proc. Natl. Acad. Sci. USA* **99**:7986-7991, 2002 ("Ottschytsch"; copy of the manuscript provided in **Exhibit E**); and Vega-Saenz de Miera, *Brain Res. Mol. Brain Res.* **123**:91-103, 2004 ("Vega-Saenz"; copy of the abstract provided in **Exhibit F**)), thus confirming Applicants' assertion that the presently claimed sequence, which is identical to the Kv10.1a and Kv6.3 proteins described above, is a voltage-gated potassium ion channel protein. Applicants respectfully point out that whether or not the Sano, Ottschytsch, and Vega-Saenz references cited by Applicants above were available at the time of filing of the present application is not germane to the utility issue at hand. Applicants point to the Sano, Ottschytsch, and Vega-Saenz references **not** to evidence that these sequences were known in the art at the time the present application was filed, but, rather, to evidence that other skilled artisans have **confirmed** Applicants' assertion that the presently claimed sequence is a voltage-gated potassium ion channel protein.

The Examiner states that "the specification does not disclose disorders or conditions associated with a mutated, deleted, or translocated gene" (the Action at page 7). First, Applicants point out that the disclosure of "disorders or conditions associated with a mutated, deleted, or translocated gene" is not the standard for patentability under 35 U.S.C. § 101 (*In re Brana*, 34 USPQ2d 1436 (Fed. Cir. 1995); "*Brana*"). Second, and more importantly, Applicants respectfully point out that the presently claimed sequence has clearly been described by Applicants in the specification as originally filed as a voltage-gated ion channel protein (see, at least, page 2, line 5 of the specification) that is involved in

high blood pressure, arrhythmia, and diabetes (see, at least, page 13, lines 19-22 of the specification). Furthermore, Applicants respectfully point out that present sequence has been shown to specifically modulate Kv2.1 voltage-gated potassium ion channel subunits (see **Exhibits D-F**), and that the association of Kv2.1 voltage-gated potassium ion channel subunits and high blood pressure (Michelakis *et al.*, *Adv. Exp. Med. Biol.* **502**:401-418, 2001; copy of abstract provided in **Exhibit G**), arrhythmia (Lee *et al.*, *Am. J. Physiol.* **277**:H1725-H1731, 1999 (copy of article provided in **Exhibit H**) and Huang *et al.*, *J. Cardiovasc. Electrophysiol.* **11**:1252-1261, 2000 (copy of abstract provided in **Exhibit I**)), and diabetes (MacDonald *et al.*, *Mol. Endocrinol.* **15**:1423-1435, 2001 (copy of article provided in **Exhibit J**) and Qin *et al.*, *Biochem. Biophys. Res. Commun.* **283**:549-553, 2001 (copy of abstract provided in **Exhibit K**)) were all well-known in the art at the time the present application was filed. Example 10 of the Revised Interim Utility Guidelines Training Materials (pages 53-55; **Exhibit L**), which have been set forth by the United States Patent and Trademark Office ("the USPTO"), clearly establishes that a rejection under 35 U.S.C. § 101 as allegedly lacking a patentable utility, and under 35 U.S.C. § 112, first paragraph, as allegedly unusable by the skilled artisan due to the alleged lack of patentable utility (see Section VI, below), is not proper when a full length sequence (such as the presently claimed sequence) has a similarity score greater than 95% to a protein having a well-established utility. Therefore, based on the 100% identity between the presently claimed sequence and Kv6.3 and Kv10.1a, the established interaction between the claimed sequence and Kv2.1, and the association of Kv2.1 and high blood pressure, arrhythmia and diabetes, as detailed above, as the present situation **exactly** tracks Example 10 of the Revised Interim Utility Guidelines Training Materials, the USPTO's **own** examination guidelines clearly indicate that the present claims meet the requirements of 35 U.S.C. § 101 and 35 U.S.C. § 112, first paragraph (see Section VI, below). Thus, the present rejection of claims 3, 5, 6, 8, 9 and 11 should be withdrawn.

The Examiner cites a number of articles (Wells, *Biochemistry* **29**:8509-8517, 1990; Ngo *et al.*, "The Protein Folding Problem and Tertiary Structure Prediction", pp. 492-495, 1994; Lehman-Horn *et al.*, *Physiol. Rev.* **79**:1317-1372, 1999; Bork, *Genome Res.* **10**:398-400, 2000, Skolnick *et al.*, *Trends in Biotech.* **18**:34-39, 2000; Doerks *et al.*, *Trends in Genetics* **14**:248-250, 1998; Smith *et al.*, *Nature Biotechnology* **15**:1222-1223, 1997; Brenner, *Trends in Genetics* **15**:132-133, 1999; and Bork *et al.*, *Trends in Genetics* **12**:425-427, 1996) for the proposition that

"(t)he problem of predicting protein and DNA structure from sequence data and in turn utilizing predicted structural determinations to ascertain functional aspects of the protein and DNA is extremely complex" (the Action at page 8), and, thus, "the art recognizes that function cannot be predicted from structure alone" (the Action at page 9). Applicants suggest that such citations reflect that the Examiner appears to believe that extensive structural similarity is not enough to establish a specific utility. First, Applicants respectfully point out the Examiner's position **directly contradicts** the position of the USPTO itself, as set forth in Example 10 of the Revised Interim Utility Guidelines Training Materials (see **Exhibit L**), which clearly establishes that structural similarity **can** in fact be used to establish function, and thus establish a specific utility. Second, rather than detail the numerous failings of each of the articles cited by the Examiner, Applicants merely note for the record that scientific manuscripts from as far back as 1990 can hardly be considered to reflect the state of the art at the time the present application was filed. Therefore, as the USPTO's **own** examination guidelines clearly indicate that structural similarity **can** in fact be used to establish function, and thus establish a specific utility, the present claims meet the requirements of 35 U.S.C. § 101 and 35 U.S.C. § 112, first paragraph (see Section VI, below). Thus, the present rejection of claims 3, 5, 6, 8, 9 and 11 should be withdrawn.

It has been well established that Applicants need only make one credible assertion of utility to meet the requirements of 35 U.S.C. § 101 (*Raytheon v. Roper*, 220 USPQ 592 (Fed. Cir. 1983); *In re Gottlieb*, 140 USPQ 665 (CCPA 1964); *In re Malachowski*, 189 USPQ 432 (CCPA 1976); *Hoffman v. Klaus*, 9 USPQ2d 1657 (Bd. Pat. App. & Inter. 1988)), and, thus, any questions concerning whether or not the present claims meet the requirements of 35 U.S.C. § 101 should have been laid to rest. Nevertheless, Applicants respectfully point out that the present invention has a number of additional substantial and credible utilities, not the least of which is in forensic biology, as described in the specification as originally filed, at least at page 3, line 12. As described in the specification as originally filed, at page 17, lines 3-5, a coding single nucleotide polymorphism was identified in the presently claimed sequence - specifically, a silent G/C polymorphism at nucleotide position 432 of SEQ ID NO:1, both of which result in a glycine being present at amino acid position 144 of SEQ ID NO:2. As such polymorphisms are the basis for forensic analysis, which is undoubtedly a "real world" utility, the presently claimed sequence must in itself be useful. Thus, the present claims clearly meet the requirements of 35 U.S.C. § 101.

Applicants respectfully point out that, even though Applicants asserted in the specification as originally filed that the presently claimed sequence is involved in high blood pressure, arrhythmia, and diabetes (see above), the use of the presently described polymorphism in forensic analysis does **not even require** the identification of a specific medical condition. One aspect of forensic analysis is to distinguish individual members of the human population from one another based solely on the **presence** or **absence** of one or more polymorphic markers, such as the presently described polymorphism. As polymorphic markers such as the presently described polymorphism have been used in forensic analysis for decades, this is clearly a well-established technique, and as such, specific guidance does not need to be provided in the present specification, for it has long been established that a patent need not disclose what is well-known in the art (*In re Wands*, 8 USPQ 2d 1400 (Fed. Cir. 1988)). Thus, the Examiner's argument does not support the alleged lack of utility.

This is also not a case of a "potential" utility. Using the polymorphic marker exactly as described in the specification as originally filed, the skilled artisan can readily distinguish individuals from one another. Applicants respectfully point out that while using forensic analysis to make a positive identification would require information concerning the percentage of a population that contains the polymorphism, **elimination** of an individual from a pool of suspects requires no information at all concerning the percentage of a population that contains the polymorphism. Applicants point out that in the worst case scenario, each polymorphic marker is useful to eliminate 50% of the population (in other words, the marker being present in half of the population). This is an inherent feature of any polymorphic marker, as the largest percentage of a population that two polymorphic markers can define is 50% each. If a polymorphic marker is present at a level of less than 50%, then that marker is even **more** informative, *i.e.*, a **greater** percentage of the population can be eliminated on the basis of the marker. Nevertheless, the ability to eliminate even 50% of the population from a pool of suspects clearly is a real world, practical utility.

Applicants point out that naturally occurring genetic polymorphisms such as the polymorphism described in the specification as originally filed are both the basis of, and critical to, *inter alia*, forensic genetic analysis intended to resolve issues of, for example, identity or paternity. Forensic analysis based on polymorphisms such as the polymorphism identified by Applicants is used to rule out suspects in many criminal cases, and to rule out suspects in the identification of human remains. Paternity

determination is based on polymorphisms such as the polymorphism identified by Applicants to rule out individuals suspected of fathering a particular child. What could be possibly be more substantial and real world than the loss of an individual's freedom or life through incarceration? What could be possibly be more substantial and real world than the identification of human remains? What could be possibly be more substantial and real world than the impact, both economic and emotional, that the results of a paternity analysis has on the individuals directly and indirectly involved? These are all well known and generally accepted uses of polymorphisms such as the polymorphism identified by Applicants. Without such identified polymorphisms, the skilled artisan would not be able to carry out such forensic or paternal analyses. Therefore, as the use of the presently described polymorphic marker in forensic analysis is clearly a real world and substantial utility, the presently claimed sequences meet the requirements of 35 U.S.C. § 101.

The Examiner seems to imply that the present claims do not meet the requirements of 35 U.S.C. § 101 because "extensive experimentation" (the Action at page 3), "significant further research" (the Action at page 4), "(s)ignificant further experimentation" (the Action at page 6 (twice) and page 7) would be required in certain aspects of the invention. Applicants first point out that the use of the presently described polymorphic marker in forensic analysis, as detailed above, requires **no** further research. Thus, the presently described polymorphism can be used to eliminate an individual from a pool of suspects in its currently available form. Second, Applicants respectfully point out that the proper standard for meeting the requirements of 35 U.S.C. § 101 is **not** whether "extensive experimentation", "significant further research", or "(s)ignificant further experimentation" is required to practice certain aspects of the claimed invention, but whether **undue** experimentation would be required to practice the claimed invention. The widespread use of polymorphisms such as that described by Applicants in forensic analysis every day **strongly** argues against such a use requiring "undue experimentation". Applicants point out that in assessing the question of whether undue experimentation would be required in order to practice the claimed invention, the key term is "undue", not "experimentation". *In re Angstadt and Griffin*, 190 USPQ 214 (CCPA 1976). However, even if, *arguendo*, further research might be required in certain aspects of the present invention, this does not preclude a finding that the invention has utility. As clearly set forth by the Federal Circuit in *In re Brana,* (34 USPQ2d 1436 (Fed. Cir. 1995); "*Brana*"), "pharmaceutical inventions, necessarily

includes the expectation of further research and development" (*Brana* at 1442-1443, emphasis added). Thus, the need for some experimentation clearly does not render the claimed invention unpatentable (see also *In re Wands, supra*). Indeed, a considerable amount of experimentation may be permissible if such experimentation is routinely practiced in the art. *In re Angstadt and Griffin, supra*; *Amgen, Inc. v. Chugai Pharmaceutical Co., Ltd.*, 18 USPQ2d 1016 (Fed. Cir. 1991). Thus, the present claims clearly meet the requirements of 35 U.S.C. § 101.

Applicants respectfully point out that as the presently described polymorphism is a part of the family of polymorphisms that have a well-established utility, the Federal Circuit's holding in *Brana* (*supra*) is directly on point. In *Brana*, the Federal Circuit admonished the USPTO for confusing "the requirements under the law for obtaining a patent with the requirements for obtaining government approval to market a particular drug for human consumption". *Brana* at 1442. The Federal Circuit went on to state:

> At issue in this case is an important question of the legal constraints on patent office examination practice and policy. The question is, with regard to pharmaceutical inventions, what must the applicant provide regarding the practical utility or usefulness of the invention for which patent protection is sought. This is not a new issue; it is one which we would have thought had been settled by case law years ago.

Brana at 1439, emphasis added. The choice of the phrase "utility or usefulness" in the foregoing quotation is highly pertinent. The Federal Circuit is evidently using "utility" to refer to rejections under 35 U.S.C. § 101, and is using "usefulness" to refer to rejections under 35 U.S.C. § 112, first paragraph. This is made evident in the continuing text in *Brana*, which explains the correlation between 35 U.S.C. §§ 101 and 112, first paragraph. The Federal Circuit concluded:

> FDA approval, however, is not a prerequisite for finding a compound useful within the meaning of the patent laws. Usefulness in patent law, and in particular in the context of pharmaceutical inventions, necessarily includes the expectation of further research and development. The stage at which an invention in this field becomes useful is well before it is ready to be administered to humans. Were we to require Phase II testing in order to prove utility, the associated costs would prevent many companies from obtaining patent protection on promising new inventions, thereby eliminating an incentive to pursue, through research and development, potential cures in many crucial areas such as the treatment of cancer.

Brana at 1442-1443, citations omitted. Thus, based on the holding in *Brana*, the present claims meet the requirements under 35 U.S.C. § 101 and 35 U.S.C. § 112, first paragraph (see Section VI,

below).

It is important to note that it has been clearly established that a statement of utility in a specification must be accepted absent reasons why one skilled in the art would have reason to doubt the objective truth of such statement. *In re Langer*, 503 F.2d 1380, 1391, 183 USPQ 288, 297 (CCPA, 1974; "*Langer*"); *In re Marzocchi*, 439 F.2d 220, 224, 169 USPQ 367, 370 (CCPA, 1971). As clearly set forth in *Langer*:

> As a matter of Patent Office practice, a specification which contains a disclosure of utility which corresponds in scope to the subject matter sought to be patented must be taken as sufficient to satisfy the utility requirement of § 101 for the entire claimed subject matter unless there is a reason for one skilled in the art to question the objective truth of the statement of utility or its scope.

Langer at 297, emphasis in original. As set forth in the Manual of Patent Examining Procedure ("MPEP"), "Office personnel must provide evidence sufficient to show that the statement of asserted utility would be considered 'false' by a person of ordinary skill in the art" (MPEP, Eighth Edition at 2100-40, emphasis added). Therefore, absent evidence from the Examiner that the presently described polymorphic marker could not be used in forensic analysis as detailed above, as the skilled artisan would readily understand that the present polymorphic marker has utility in forensic analysis, the present claims clearly meet the requirements of 35 U.S.C. § 101.

Additionally, given the association between the presently claimed sequence and high blood pressure, arrhythmia, and diabetes, as detailed above, those of skill in the art would readily appreciate the importance of tracking the expression of the genes encoding the described proteins, as described in the specification as originally filed, at least at page 6, lines 5-7. In particular, the specification describes how the described sequences can be represented using a gene chip format to provide a high throughput analysis of the level of gene expression. Such "DNA chips" clearly have utility, as evidenced by hundreds of issued U.S. Patents, as exemplified by U.S. Patent Nos. 5,445,934, 5,556,752, 5,744,305 (**Exhibits M-O**; submitted with the Information Disclosure Statement filed on March 5, 2002), and U.S. Patent Nos. 5,837,832, 6,156,501 and 6,261,776 (**Exhibits P-R**; copies of issued U.S. Patents not provided pursuant to requests from the USPTO). As the present sequences are specific markers of the human genome (see below), and such specific markers are targets for the discovery of drugs that are associated with human disease, those of skill in the art would instantly

recognize that the present nucleotide sequences would be an ideal, novel candidate for assessing gene expression using such DNA chips. Given the widespread utility of such "gene chip" methods using *public domain* gene sequence information, there can be little doubt that the use of the presently described *novel* sequences would have great utility in such DNA chip applications. Clearly, compositions that enhance the utility of such DNA chips, such as the presently claimed nucleotide sequences, must in themselves be useful.

Further evidence of the "real world" substantial utility of the present invention is provided by the fact that there is an entire industry established based on the use of gene sequences or fragments thereof in a gene chip format. Perhaps the most notable gene chip company is Affymetrix. However, there are many companies which have, at one time or another, concentrated on the use of gene sequences or fragments, in gene chip and non-gene chip formats, for example: Gene Logic, ABI-Perkin-Elmer, HySeq and Incyte. In addition, one such company (Rosetta Inpharmatics) was viewed to have such "real world" value that it was acquired by large a pharmaceutical company (Merck) for significant sums of money (net equity value of the transaction was $620 million). The "real world" substantial industrial utility of gene sequences or fragments would, therefore, appear to be widespread and well established. Clearly, persons of skill in the art, as well as venture capitalists and investors, readily recognize the utility, both scientific and commercial, of genomic data in general, and specifically human genomic data. Billions of dollars have been invested in the human genome project, resulting in useful genomic data (see, *e.g.*, Venter *et al.*, *Science* **291**:1304, 2001; **Exhibit S**). The results have been a stunning success as the utility of human genomic data has been widely recognized as a great gift to humanity (see, *e.g.*, Jasny and Kennedy, *Science* **291**:1153, 2001; **Exhibit T**). Clearly, the usefulness of human genomic data, such as the presently claimed nucleic acid molecules, is substantial and credible (worthy of billions of dollars and the creation of numerous companies focused on such information) and well-established (the utility of human genomic information has been clearly understood for many years). Thus, the present claims clearly meet the requirements of 35 U.S.C. § 101.

The Examiner alleges that this asserted utility is "not specific or substantial" because "any polynucleotide sequence" can be used in this manner (the Action at page 6). This argument is flawed in a number of respects. First, Applicants respectfully point out that the association between the presently claimed sequence and high blood pressure, arrhythmia, and diabetes, as detailed above, is

not true for "any polynucleotide sequence". Furthermore, expression profiling does not even require a knowledge of the function of the particular nucleic acid on the chip - rather the gene chip indicates which DNA fragments are expressed at greater or lesser levels in two or more particular tissue types. Skilled artisans already have used and continue to use sequences such as Applicants in gene chip applications without further experimentation. Second, Applicants respectfully point out that only expressed polynucleotide sequences can be used to track gene expression, not just "any polynucleotide sequence". Third, the Examiner appears to be confusing the requirement for a specific utility, which is the proper standard for utility under 35 U.S.C. § 101, with a requirement for a unique utility, which is clearly an improper standard. As clearly set forth by the Federal Circuit in *Carl Zeiss Stiftung v. Renishaw PLC*, 20 USPQ2d 1101 (Fed. Cir. 1991; "*Carl Zeiss*"):

> An invention need not be the best or only way to accomplish a certain result, and it need only be useful to some extent and in certain applications: "[T]he fact that an invention has only limited utility and is only operable in certain applications is not grounds for finding a lack of utility." *Envirotech Corp. v. Al George, Inc.*, 221 USPQ 473, 480 (Fed. Cir. 1984)

Following directly from the quote above, an invention does not need to be the only way to accomplish a certain result. Thus, the question of whether or not other nucleic acid sequences can be used to assess gene expression patterns is completely irrelevant to the present utility inquiry. The only relevant question in regard to meeting the standards of 35 U.S.C. § 101 is whether "any polynucleotide sequence" can be so used - and the clear answer to this question is an emphatic **no**. Importantly, the holding in *Carl Zeiss* is mandatory legal authority that essentially controls the outcome of the present case. This case, and particularly the cited quote, directly rebuts the Examiner's argument. Furthermore, the requirement for a unique utility is clearly not the standard adopted by the USPTO. If every invention were required to have a unique utility, the USPTO would no longer be issuing patents on batteries, automobile tires, golf balls, golf clubs, and treatments for a variety of human diseases, such as cancer and bacterial or viral infections, just to name a few particular examples, because examples of each of these have already been described and patented. All batteries have the exact same utility - specifically, to provide power. All automobile tires have the exact same utility - specifically, for use on automobiles. All golf balls and golf clubs have the exact same utility - specifically, use in the game of golf. All cancer treatments have the exact same utility - specifically, to treat cancer. All anti-infectious

agents have the exact same broader utility - specifically, to treat infections. However, only the briefest perusal of virtually any issue of the Official Gazette provides numerous examples of patents being granted on each of the above compositions every week. Additionally, if a composition needed to be unique to be patented, the entire class and subclass system would be an effort in futility, as the class and subclass system serves solely to group such common inventions, which would not be required if each invention needed to have a unique utility. Thus, the present sequence clearly meets the requirements of 35 U.S.C. § 101.

Applicants note that the Examiner correctly determines that the generic class with regard to the present invention is "any polynucleotide sequence", but then attempts to narrow the generic class of the invention to include only those nucleic acids that are expressed (and associated with high blood pressure, arrhythmia, and diabetes) in order to support an allegation that the claimed nucleic acids lack a "specific" utility. Applicants reiterate that not all nucleic acids are expressed - in fact, only 2-4% of all nucleotide sequences are expressed, and only a very small number of these are associated with high blood pressure, arrhythmia, and diabetes. Therefore, the question of whether the asserted utility is "specific", as opposed to "generic", has clearly been laid to rest. Applicants note that such redefinition of the generic class of the invention is completely improper, and in clear defiance of established case law. Therefore the present claims are clearly in compliance with 35 U.S.C. § 101.

The Examiner further discounts this assertion of utility because "the specification does not disclose specific cDNA ... targets" (the Action at page 6). This is simply not true. The specification as originally filed at page 3, lines 28-33, clearly states that the presently claimed sequence "is expressed in human fetal brain, brain, cerebellum, pituitary, prostate, thymus, lymph node, bone marrow, trachea, fetal liver, liver, testis, thyroid, salivary gland, stomach, skeletal muscle, heart, uterus, adipose, hypothalamus, ovary, tongue, aorta, 12 week old embryo, adenocarcinoma, and osteosarcoma cells". Thus, the Examiner's argument in no way supports the allegation that the presently claimed sequences lack a patentable utility.

As yet a further example of the utility of the presently claimed polynucleotides, as described in the specification at page 3, lines 2-4, the present nucleotide sequences have a specific utility in "the identification of protein coding sequences" and "mapping a unique gene to a particular chromosome". The specification as originally filed, at page 3, lines 5 and 6, details that the gene encoding the presently

claimed sequences is present on "human chromosome 2, see GENBANK accession no. AC0025750". In fact, alignment of SEQ ID NO:1 with GenBank Accession Number AC025750 (which is a genomic clone from human chromosome 2) shows that the human gene corresponding to SEQ ID NO:1 is dispersed on 2 exons of human chromosome 2 (alignment and the first page from the GenBank report are presented in **Exhibit U**). Clearly, the present polynucleotide provides exquisite specificity in localizing the specific region of human chromosome 2 that contains the gene encoding the given polynucleotide, a utility not shared by virtually any other nucleic acid sequences. In fact, it is this specificity that makes this particular sequence so useful. Early gene mapping techniques relied on methods such as Giemsa staining to identify regions of chromosomes. However, such techniques produced genetic maps with a resolution of only 5 to 10 megabases, far too low to be of much help in identifying specific genes involved in disease. The skilled artisan readily appreciates the significant benefit afforded by markers that map a specific locus of the human genome, such as the present nucleic acid sequence. For further evidence in support of the Applicants' position, the Examiner is requested to review, for example, section 3 of Venter *et al.* (*supra*, at pp. 1317-1321, including Fig. 11 at pp.1324-1325; see **Exhibit S**), which demonstrates the significance of expressed sequence information in the structural analysis of genomic data. The presently claimed polynucleotide sequence defines a biologically validated sequence that provides a unique and specific resource for mapping the genome essentially as described in the Venter *et al.* article. Thus, the present claims clearly meet the requirements of 35 U.S.C. § 101.

Applicants reiterate that only a minor percentage (2-4%) of the genome actually encodes exons, which in-turn encode amino acid sequences. Significantly, the claimed polynucleotide sequence defines how the encoded exons are actually spliced together to produce an active transcript (*i.e.*, the described sequences are useful for functionally defining exon splice-junctions). As described in the specification as originally filed at page 3, lines 6-9, the claimed "sequences identify actual, biologically relevant, exon splice junctions, as opposed to those that might have been predicted bioinformatically from genomic sequence alone". The specification as originally filed, at page 11, lines 13-18, further details that "sequences derived from regions adjacent to the intron/exon boundaries of the human gene can be used to design primers for use in amplification assays to detect mutations within the exons, introns, splice sites (*e.g.*, splice acceptor and/or donor sites), *etc.*, that can be used in diagnostics and

pharmacogenomics". Applicants respectfully submit that the practical scientific value of biologically validated, expressed, spliced, and polyadenylated mRNA sequences is readily apparent to those skilled in the relevant biological and biochemical arts. Thus, the present sequence clearly meets the requirements of 35 U.S.C. § 101.

Once again, the Examiner alleges that this asserted utility is "not specific or substantial" because "(s)uch assays can be performed with any polynucleotide" (the Action at page 7). With respect to the presently asserted utility, this argument is once again flawed in a number of respects. First, Applicants once again point out that only expressed sequences can be used in the identification of coding sequence, not just "any polynucleotide". Second, Applicants reiterate that the requirements of a specific utility, which is the proper standard for utility under 35 U.S.C. § 101, should not be confused with the requirement for a unique utility, which is clearly an improper standard (*Carl Zeiss, supra*). The fact that a small number of other nucleotide sequences could be used to map the protein coding regions in this specific region of chromosome 2 does not mean that the use of Applicants' sequence to map the protein coding regions of chromosome 2 is not a specific utility. Once again, the question of whether or not other nucleic acid sequences can be so used is completely irrelevant to the present utility inquiry. The only relevant question in regard to meeting the standards of 35 U.S.C. § 101 is whether "any polynucleotide" can be so used - and the clear answer to this question is once again an emphatic **no**. Applicants respectfully point out the Examiner is once again attempting to narrow the generic class of "any polynucleotide" to include only the small number of nucleic acid molecules that are expressed from this particular region of chromosome 2 in order to support the allegation that the claimed nucleic acids lack a "specific" utility. Applicants respectfully point out once again that this is improper under the law as well as the policy of the USPTO. Thus, the present claims clearly meet the requirements of 35 U.S.C. § 101.

Rather, as set forth by the Federal Circuit, "(t)he threshold of utility is not high: An invention is 'useful' under section 101 if it is capable of providing some identifiable benefit." *Juicy Whip Inc. v. Orange Bang Inc.*, 51 USPQ2d 1700 (Fed. Cir. 1999) (citing *Brenner v. Manson*, 383 U.S. 519, 534 (1966)). Additionally, the Federal Circuit has stated that "(t)o violate § 101 the claimed device must be totally incapable of achieving a useful result." *Brooktree Corp. v. Advanced Micro Devices, Inc.*, 977 F.2d 1555, 1571 (Fed. Cir. 1992), emphasis added. *Cross v. Iizuka* (224 USPQ 739

(Fed. Cir. 1985); "*Cross*") states "any utility of the claimed compounds is sufficient to satisfy 35 U.S.C. § 101". *Cross* at 748, emphasis added. Indeed, as discussed in Section II, above, the Federal Circuit recently emphatically confirmed that "anything under the sun that is made by man" is patentable (*State Street Bank & Trust Co. v. Signature Financial Group Inc.*, *supra*, citing the U.S. Supreme Court's decision in *Diamond vs. Chakrabarty*, *supra*).

Finally, the requirements set forth in the Action for compliance with 35 U.S.C. § 101 do not comply with the requirements set forth by the USPTO itself for compliance with 35 U.S.C. § 101. While Applicants are well aware of the new Utility Guidelines set forth by the USPTO, Applicants respectfully point out that the current rules and regulations regarding the examination of patent applications is and always has been the patent laws as set forth in 35 U.S.C. and the patent rules as set forth in 37 C.F.R., not the Manual of Patent Examination Procedure or particular guidelines for patent examination set forth by the USPTO. Furthermore, it is the job of the judiciary, not the USPTO, to interpret these laws and rules. Applicants are unaware of any significant recent changes in either 35 U.S.C. § 101, or in the interpretation of 35 U.S.C. § 101 by the Supreme Court or the Federal Circuit that is in keeping with the new Utility Guidelines set forth by the USPTO. This is underscored by numerous patents that have been issued over the years that claim nucleic acid fragments that do not comply with the new Utility Guidelines. As just a few examples of such issued U.S. Patents, the Examiner is invited to review U.S. Patent Nos. 5,817,479, 5,654,173, and 5,552,281 (each of which claims short polynucleotides; **Exhibits V-X**; copies of issued U.S. Patents not provided pursuant to requests from the USPTO), and U.S. Patent No. 6,340,583 (which includes no working examples; **Exhibit Y**; copies of issued U.S. Patents not provided pursuant to requests from the USPTO), none of which contain examples of the "real-world" utilities that the Examiner seems to be requiring. As issued U.S. Patents are presumed to meet all of the requirements for patentability, including 35 U.S.C. §§ 101 and 112, first paragraph (see Section VI, below), Applicants submit that the present polynucleotides must also meet the requirements of 35 U.S.C. § 101. While Applicants understand that each application is examined on its own merits, Applicants are unaware of any changes to 35 U.S.C. § 101, or in the interpretation of 35 U.S.C. § 101 by the Supreme Court or the Federal Circuit, since the issuance of these patents that render the subject matter claimed in these patents, which is similar to the subject matter in question in the present application, as suddenly non-statutory or failing

to meet the requirements of 35 U.S.C. § 101. Thus, holding Applicants to a different standard of utility would be arbitrary and capricious, and, like other clear violations of due process, cannot stand.

For each of the foregoing reasons, Applicants submit that as the presently claimed nucleic acid molecules have been shown to have a substantial, specific, credible and well-established utility, the rejection of claims 1, 3, and 5-11 under 35 U.S.C. § 101 has been overcome, and request that the rejection be withdrawn.

VI. Rejection of Claims 1, 3, and 5-11 Under 35 U.S.C. § 112, First Paragraph

The Action next rejects claims 1, 3, and 5-11 under 35 U.S.C. § 112, first paragraph, since allegedly one skilled in the art would not know how to use the invention, as the invention allegedly is not supported by a specific, substantial, and credible utility or a well-established utility. Applicants respectfully traverse.

First, while Applicants in no way agree with the Examiner's position that one skilled in the art would not know how to use the invention as set forth in claims 1, 7 and 10, since claims 1, 7 and 10 have been cancelled entirely without prejudice and without disclaimer, the present rejection of claims 1, 7 and 10 under 35 U.S.C. § 112, first paragraph is rendered moot. The remainder of this section will therefore focus on claims 3, 5, 6, 8, 9 and 11.

Applicants submit that as claims 3, 5, 6, 8, 9 and 11 have been shown to have "a specific, substantial, and credible utility", as detailed in section V above, the present rejection of claims 3, 5, 6, 8, 9 and 11 under 35 U.S.C. § 112, first paragraph, cannot stand.

Applicants therefore request that the rejection of claims 1, 3, and 5-11 under 35 U.S.C. § 112, first paragraph, be withdrawn.

VII. Rejection of Claims 1, 6, and 9 Under 35 U.S.C. § 112, First Paragraph

The Action next rejects claims 1, 6, and 9 under 35 U.S.C. § 112, first paragraph, as allegedly not providing enablement for the full scope of the claimed invention. While Applicants in no way agree with the Examiner's position that claims 1, 6, and 9 are not enabled for the full scope of the claim, as claim 1 has been cancelled entirely without prejudice and without disclaimer, and claim 6 has been amended to reference claim 3, which is not subject to the present rejection, the present rejection of

claims 1, 6, and 9 under 35 U.S.C. § 112, first paragraph, has been overcome.

Applicants therefore respectfully request that the rejection of claims 1, 6, and 9 under 35 U.S.C. § 112, first paragraph, be withdrawn.

VIII. Rejection of Claims 9-11 Under 35 U.S.C. § 112, First Paragraph

The Action next rejects claims 9-11 under 35 U.S.C. § 112, first paragraph, as allegedly not enabled for the full scope of the claim. Applicants respectfully traverse.

First, while Applicants in no way agree with the Examiner's position that the invention as set forth in claim 10 is not fully enabled, since claim 10 has been cancelled entirely without prejudice and without disclaimer, the present rejection of claim 10 under 35 U.S.C. § 112, first paragraph is rendered moot. The remainder of this section will therefore focus on claims 9 and 11.

The Examiner states that "(t)he specification of the instant application teaches that NHP gene products (SEQ ID NO:1) can be expressed in transgenic animals and any technique known in the art may be used to introduce a NHP transgene into animals to produce the founder lines of transgenic animals", but that claims 9 and 11 are not enabled for non-human transgenic animals because "there are no methods or working examples disclosed in the instant application whereby a multicellular animal with the incorporated NHP gene of SEQ ID NO:1 is demonstrated to express the NHP peptide", and "(t)he unpredictability of the art is *very high* with regards to making transgenic animals" (the Action at page 10, emphasis in original). With regard to the Examiner's first argument, Applicants respectfully point out that this argument is not dispositive as to the question of enablement, for it has long been established that "there is no statutory requirement for the disclosure of a specific example" (*In re Gay*, 309 F.2d 769, 135 USPQ 311 (CCPA, 1962)). Thus, this argument alone cannot support an allegation that claims 9 and 11 are not enabled.

With regard to the Examiner's second argument, concerning the unpredictability in the art with regard to making transgenic animals, the Examiner cites four scientific articles that allegedly support this position. Specifically, the Examiner cites Wang *et al.* (*Nuc. Acids Res.* **27**:4609-4618, 1999) and Kaufman *et al.* (*Blood* **94**:3178-3184, 1999) to support the argument that expression levels of an inserted transgene are highly variable, Wigley *et al.* (*Reprod. Fert. Dev.* **6**:585-588, 1994) to support the argument that production of non-human transgenic animals by pronuclear microinjection (one of the

methods of producing non-human transgenic animals specifically cited in the specification as originally filed) suffers from limitations such as low frequency of integration events and random integration, and Campbell *et al.* (*Theriology* **47**:63-72, 1997) to support the argument that production of non-human transgenic animals by from ES cells (another method of producing non-human transgenic animals specifically cited in the specification as originally filed) has been difficult. Thus, the Examiner concludes that "it would have required undue experimentation for the skilled artisan to have made any and all transgenic non-human animals according to the instant invention" (the Action at page 11).

Rather than list the numerous deficiencies of each of the articles cited by the Examiner, Applicants instead will present evidence of the state of the art with regard to making transgenic animals as of the filing date of the present application (December 10, 2001). The specification as originally filed, at page 17, lines 15-18, details that "(a)nimals of any species, including, but not limited to, worms, mice, rats, rabbits, guinea pigs, pigs, micro-pigs, birds, goats, and non-human primates, *e.g.*, baboons, monkeys, and chimpanzees may be used to generate NHP transgenic animals". Applicants respectfully point out that there are **numerous** examples of transgenic worms (nematodes), mice, rats, rabbits, guinea pigs, pigs, birds (chickens), goats and monkeys, years and sometimes decades prior to the filing date of the present application. However, rather than provide hundreds of citations of transgenic animals that are in the art prior to the filing date of the present application, Applicants respectfully point out that the first report of a transgenic nematode was in 1988 (Spieth *et al.*, *Dev. Biol.* **130**:285-293; copy of abstract provided in **Exhibit Z**), the first report of a transgenic mouse was in 1980 (Gordon *et al.*, *Proc. Natl. Acad. Sci. USA* **77**:7380-7384; copy of manuscript provided in **Exhibit AA**), the first report of a transgenic rat was in 1990 (Mullins *et al.*, *Nature* **344**:541-544; copy of abstract provided in **Exhibit BB**), the first report of a transgenic rabbit was in 1985 (Hammer *et al.*, *Nature* **315**:680-683; copy of abstract provided in **Exhibit CC**), a report of the production of human interleukin-2 in the milk of transgenic rabbits was published in 1990 (Bühler *et al.*, *Bio/Technology* **8**:140-143; copy of abstract provided in **Exhibit DD**), the first reports of transgenic guinea pigs were in 2000 (Suzuki *et al.*, *Gene Ther.* **7**:1046-1054, and Yagi *et al.*, *JARO* **1**:315-325; copies of abstracts provided in **Exhibit EE**), a report of the production of human growth hormone in the milk of transgenic guinea pigs was also published in 2000 (Hens *et al.*, *Biochim. Biophys. Acta* **1523**:161-171; copy of abstract provided in **Exhibit FF**), the first report of a transgenic pig was in

1985 (see **Exhibit CC**), a report of the production of a heterologous milk protein in the milk of transgenic pigs was published in 1991 (Wall *et al.*, *Proc. Natl. Acad. Sci. USA* **88**:1696-1700; copy of manuscript provided in **Exhibit GG**), the first reports of transgenic chickens were in 1987 (Salter *et al.*, *Virology* **157**:236-240; copy of abstract provided in **Exhibit HH**) and 1989 (Bosselman *et al.*, *J. Virol.* **63**:2680-2689; copy of abstract provided in **Exhibit II**), the first reports of transgenic goats were in 1991 (Ebert *et al.*, *Bio/Technology* **9**:835-838, and Denman *et al.*, *Bio/Technology* **9**:839-843; copies of abstracts provided in **Exhibit JJ**), and the first report of a transgenic monkey (rhesus monkey) was in January of 2001 (Chan *et al.*, *Science* **291**:309-312; copy of manuscript provided in **Exhibit KK**). Additionally, the first report of a transgenic cow (raised by the Examiner on page 11 of the Action) was in 1991 (Krimpenfort *et al.*, *Bio/Technology* **9**:844-847; copy of abstract provided in **Exhibit LL**), the first report of a transgenic sheep (another example of a transgenic mammal) was in 1988 (Simons et al., *Bio/Technology* **6**:179-183; copy of abstract provided in **Exhibit MM**), and a report of the production of human anti-hemophilic factor IX in the milk of transgenic sheep was published in 1989 (Clark *et al.*, *Bio/Technology* **7**:487-492; copy of abstract provided in **Exhibit NN**). Given the hundreds of reports of transgenic animals, of which the reports listed above are only the **first** examples, there can be no doubt that the making of transgenic animals is clearly enabled to those of skill in the art, which is all that is required to meet the enablement requirement under 35 U.S.C. § 112, first paragraph.

The Examiner seems to believe that claims 9 and 11 are not enabled for transgenic animals because certain aspects of transgenic technology (expression levels, site-specific *versus* random integration) require some level of experimentation to perfect. However, Applicants respectfully point out that all that is required in order to satisfy the enablement requirement under 35 U.S.C. § 112, first paragraph, is making **any** transgenic animal, not the **perfect** transgenic animal. **Any** detectable level of expression of a transgene, for example SEQ ID NO:1, is all that is required, for it is well established that the enablement requirement is met if **any** use of the invention (or in this case, certain aspects of the invention) is provided (*In re Nelson*, 126 USPQ 242 (CCPA 1960); *Cross v. Iizuka, supra*). "The enablement requirement is met if the description enables any mode of making and using the invention." *Johns Hopkins Univ. v. CellPro, Inc.*, 47 USPQ2d 1705, 1719 (Fed. Cir. 1998), citing *Engel Indus., Inc. v. Lockformer Co.*, 20 USPQ2d 1300, 1304 (Fed. Cir. 1991). Furthermore, a

specification "need describe the invention only in such detail as to enable a person skilled in the most relevant art to make and use it." *In re Naquin*, 158 USPQ 317, 319 (CCPA 1968); emphasis added. Therefore, as the skilled artisan is clearly able to make a variety of different species of transgenic animals, claims 9 and 11 are thus enabled as they are supported by a specification that provides sufficient description to enable the skilled person to make and use the invention as claimed.

The Examiner states that the present invention could not be practiced without "undue experimentation". However, it is important to remember that in assessing the question of whether undue experimentation would be required in order to practice the claimed invention, the key term is "undue", not "experimentation". *In re Angstadt and Griffin, supra*. The large number of reports in the literature on a variety of transgenic animals strongly argues against such a use requiring "undue experimentation". In *In re Wands* (*supra*), the USPTO took the position that the applicant failed to demonstrate that the disclosed biological processes of immunization and antibody selection could reproducibly result in a useful biological product (antibodies from hybridomas) within the scope of the claims. In its decision overturning the USPTO's rejection, the Federal Circuit found that Wands' demonstration of success in four out of nine cell lines screened was sufficient to support a conclusion of enablement. The court emphasized that the need for some experimentation requiring, *e.g.*, production of the biological material followed by routine screening, was not a basis for a finding of non-enablement, stating:

> Disclosure in application for the immunoassay method patent does not fail to meet enablement requirement of 35 USC 112 by requiring 'undue experimentation', even though production of monoclonal antibodies necessary to practice invention first requires production and screening of numerous antibody producing cells or 'hybridomas', since practitioners of art are prepared to screen negative hybridomas in order to find those that produce desired antibodies, since in monoclonal antibody art one 'experiment' is not simply screening of one hybridoma but rather is entire attempt to make desired antibody, and since record indicates that amount of effort needed to obtain desired antibodies is not excessive, in view of Applicants' success in each attempt to produce antibody that satisfied all claim limitations.

Wands at 1400. Thus, the need for some experimentation does not render the claimed invention unpatentable under 35 U.S.C. § 112, first paragraph. Indeed, a considerable amount of experimentation may be permissible if such experimentation is routinely practiced in the art. *In re Angstadt and Griffin, supra*; *Amgen, Inc. v. Chugai Pharmaceutical Co., Ltd., supra*). Therefore,

given the evidence detailed above concerning the ability of the skilled artisan to produce transgenic animals with some detectable level of transgene expression with reasonable certainty, claims 9 and 11 meet the enablement requirement.

The Examiner next states that claims 9 and 11 are not enabled because "(t)he specification also discloses that 'nucleotide constructs encoding such NHP products can be used to genetically engineer host cells to express such products in vivo' and that these products can be used in gene therapy" (the Action at page 12), but "the specification does not teach any methods or working examples that indicate a NHP nucleic acid is introduced and expressed in a cell for therapeutic purposes", and the "(r)elevant literature teaches that since 1990, about 3500 patients have been treated via gene therapy and although some evidence of gene transfer has been seen, it has generally been inadequate for a meaningful clinical response" (the Action at page 12), citing a journal article by Phillips (*J. Pharm. Pharmacology* **53**:1169-1174, 2001). Once again, with regard to the Examiner's first argument, Applicants respectfully point out that this argument is not dispositive as to the question of enablement, for it has long been established that "there is no statutory requirement for the disclosure of a specific example" (*In re Gay, supra*). Thus, this argument alone cannot support an allegation that claims 9 and 11 are not enabled.

With regard to the Examiner's second argument, it once again appears that the Examiner seems to believe that claims 9 and 11 are not enabled for gene therapy because gene therapy is not always effective. However, Applicants once again point out that it is well established that the enablement requirement is met if **any** use of the invention is provided (*In re Nelson, supra*; *Cross v. Iizuka, supra*). "The enablement requirement is met if the description enables any mode of making and using the invention." *Johns Hopkins Univ. v. CellPro, Inc., supra*). Furthermore, a specification "need describe the invention only in such detail as to enable a person skilled in the most relevant art to make and use it." *In re Naquin, supra*; emphasis added. Applicants respectfully point out that there are a number of reports in the literature, prior to the filing date of the present application, concerning a variety of gene therapy vectors and successful gene therapy regimens. In fact, the article by Phillips cited by the Examiner herself admits that gene therapy can and has been practiced by the skilled artisan ("since 1990, about 3500 patients have been treated via gene therapy" and "some evidence of gene transfer has been seen"). Therefore, claims 9 and 11 are clearly enabled as they are supported by a

specification that provides sufficient description to enable the skilled person to make and use the invention as claimed.

Furthermore, with regard to a requirement that the host cells of claims 9 and 11 be nearly always effective in gene therapy, such an enablement standard conflicts with established patent law. As discussed in *In re Brana* ("*Brana*", *supra*), the Federal Circuit admonished the USPTO for confusing "the requirements under the law for obtaining a patent with the requirements for obtaining government approval to market a particular drug for human consumption". Thus, based on the holding in *Brana*, claims 9 and 11 clearly meet the enablement requirement under 35 U.S.C. § 112, first paragraph.

The Examiner then once again concludes that "undue experimentation would be required of the skilled artisan to introduce and express a NHP nucleic acid into the cells of an organism" (the Action at page 12). However, Applicants reiterate that in assessing the question of whether undue experimentation would be required in order to practice the claimed invention, the key term is "undue", not "experimentation" (*In re Angstadt and Griffin, supra*). Once again, the large number of reports in the literature on a variety of gene therapy vectors, and advances in gene therapy techniques, strongly argues against such a use requiring "undue experimentation". However, even if, *arguendo*, further experimentation might be required in certain aspects of the present invention, this does not preclude a finding that the invention is enabled, as set forth by the Federal Circuit's holding in *Brana*, which clearly states, as highlighted in the quote above, that "pharmaceutical inventions, necessarily includes the expectation of further research and development" (*Brana* at 1442-1443, emphasis added). Furthermore, the need for some experimentation does not render the claimed invention unpatentable under 35 U.S.C. § 112, first paragraph (*In re Wands, supra*). Indeed, a considerable amount of experimentation may be permissible if such experimentation is routinely practiced in the art. *In re Angstadt and Griffin, supra*; *Amgen, Inc. v. Chugai Pharmaceutical Co., Ltd., supra*). Therefore, given the evidence detailed above concerning the ability of the skilled artisan to create gene therapy constructs that have some level of success, claims 9 and 11 meet the enablement requirement.

Therefore, based on the evidence of record that it is well-known to skilled artisan how to make and use a variety of species of transgenic animals, as well as a variety of gene therapy vectors, the 35 U.S.C. § 112, first paragraph, rejection is improper:

> As a matter of patent office practice, then, a specification disclosure which contains a teaching of the manner and process of making and using the invention in terms which correspond in scope to those used in describing and defining the subject matter sought to be patented must be taken as in compliance with the enabling requirement of the first paragraph of § 112 unless there is reason to doubt the objective truth of the statements contained therein which must be relied on for enabling support.

In re Marzocchi, 169 USPQ 367, 369 (CCPA 1971), emphasis as in original. Applicants respectfully point out that, as a matter of law, it is well settled that a patent need not disclose what is well-known in the art. *In re Wands, supra*. In fact, it is preferable that what is well-known in the art be omitted from the disclosure. *Hybritech, Inc. v. Monoclonal Antibodies, Inc.*, 231 USPQ 81 (Fed. Cir. 1986). Therefore, the full breadth of claims 9 and 11 are clearly enabled.

Applicants therefore request that the rejection of claims 9-11 under 35 U.S.C. § 112, first paragraph, be withdrawn.

IX. Rejection of Claims 1, 6, and 9 Under 35 U.S.C. § 112, First Paragraph

The Action next rejects claims 1, 6, and 9 under 35 U.S.C. § 112, first paragraph, as allegedly containing subject matter that was not described in the specification in such a way as to reasonably convey to one skilled in the relevant art that the inventors, at the time the application was filed, had possession of the claimed invention. While Applicants in no way agree with the Examiner's position that claims 1, 6, and 9 do not meet the requirements of 35 U.S.C. § 112, first paragraph, as claim 1 has been cancelled entirely without prejudice and without disclaimer, and claim 6 has been amended to reference claim 3, which is not subject to the present rejection, the present rejection of claims 1, 6, and 9 under 35 U.S.C. § 112, first paragraph, has been overcome.

Applicants therefore respectfully request that the rejection of claims 1, 6, and 9 under 35 U.S.C. § 112, first paragraph, be withdrawn.

X. Rejection of Claims 1, 6, and 9 Under 35 U.S.C. § 102(e)

The Action next rejects claims 1, 6, and 9 under 35 U.S.C. § 102(e), as allegedly anticipated by Tang *et al.* (US 20040014945A1; "Tang"). While Applicants do not necessarily agree with the Examiner's position that claims 1, 6, and 9 are anticipated by Tang, as claim 1 has been cancelled entirely without prejudice and without disclaimer, and claim 6 has been amended to reference claim 3,

which is not subject to the present rejection, the present rejection of claims 1, 6, and 9 under 35 U.S.C. § 102(e) has been overcome.

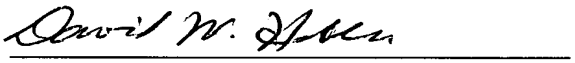
Applicants therefore respectfully request that the rejection of claims 1, 6, and 9 under 35 U.S.C. § 102(e) be withdrawn.

XI. Conclusion

The present document is a full and complete response to the Action. In conclusion, Applicants submit that, in light of the foregoing remarks, the present case is in condition for allowance, and such favorable action is respectfully requested. Should Examiner Bunner have any questions or comments, or believe that certain amendments of the claims might serve to improve their clarity, a telephone call to the undersigned Applicants' representative is earnestly solicited.

Respectfully submitted,

March 1, 2005
Date



David W. Hibler Reg. No. 41,071
Agent for Applicants

LEXICON GENETICS INCORPORATED
(281) 863-3399

Customer # 24231

EXHIBIT A

PCT WELTORGANISATION FÜR GEISTIGES EIGENTUM
Internationales Büro

INTERNATIONALE ANMELDUNG VERÖFFENTLICHT NACH DEM VERTRAG ÜBER DIE INTERNATIONALE ZUSAMMENARBEIT AUF DEM GEBIET DES PATENTWESENS (PCT)

(51) Internationale Patentklassifikation 7 :
C12N 15/12, C07K 14/705, C12Q 1/68, G01N 33/50, C07K 16/28

A2

(11) Internationale Veröffentlichungsnummer: **WO 00/08146**

(43) Internationales Veröffentlichungsdatum: 17. Februar 2000 (17.02.00)

(21) Internationales Aktenzeichen: PCT/EP99/05983

(22) Internationales Anmeldedatum: 6. August 1999 (06.08.99)

(30) Prioritätsdaten:
198 41 413.7 6. August 1998 (06.08.98) DE

(71) Anmelder *(für alle Bestimmungsstaaten ausser US)*: FORSCHUNGSGESELLSCHAFT GENION MBH [DE/DE]; Abteistrasse 57, D–20149 Hamburg (DE).

(72) Erfinder; und
(75) Erfinder/Anmelder *(nur für US)*: NETZER, Rainer [DE/DE]; Düsterntwiete 47, D–22549 Hamburg (DE). PONGS, Olaf [DE/DE]; Beim Andreasbrunnen 5, D–20249 Hamburg (DE).

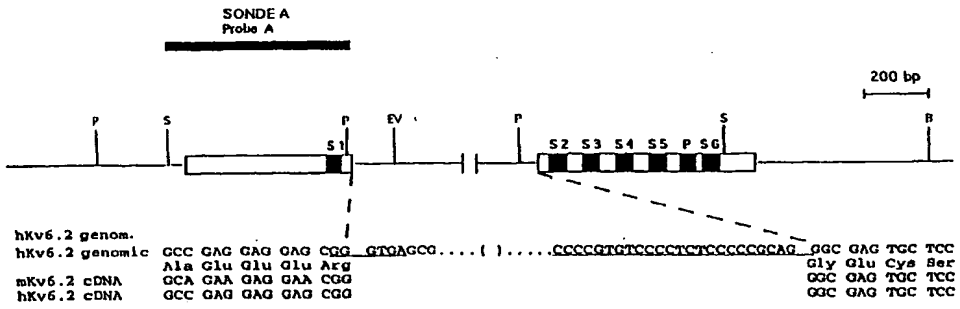
(74) Anwälte: WEBER–QUITZAU, Martin usw.; Uexküll & Stolberg, Beselerstrasse 4, D–22607 Hamburg (DE).

(81) Bestimmungsstaaten: US, europäisches Patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Veröffentlicht
Ohne internationalen Recherchenbericht und erneut zu veröffentlichen nach Erhalt des Berichts.

(54) Title: NOVEL TENSION–DEPENDENT POTASSIUM CHANNEL AND THE USE THEREOF IN THE DEVELOPMENT OF NEW THERAPEUTIC AGENTS

(54) Bezeichnung: NEUER SPANNUNGSABHÄNGIGER KALIUMKANAL UND SEINE VERWENDUNG ZUR ENTWICKLUNG VON THERAPEUTIKA



(57) Abstract

The invention relates to a novel tension–dependent potassium channel protein Kv6.2 (SEQ ID NO: 1). The Kv6.2 gene is expressed preferably in the myocardium or in the hippocampus. Novel functional heteromultimeric potassium channels having high affinity with propafenone are formed in conjunction with subunit Kv2.1. According to the invention, said novel potassium channels are used in test systems which are suitable for identifying substances modulating, opening or closing the Kv2.1/Kv6.2 channels and which can be used as therapeutic agents.

EXHIBIT B

>AF454547 ACCESSION:AF454547 NID: gi 22164081 gb AF454547.1 Homo sapiens voltage-gated potassium channel subunit Kv10.1a mRNA, complete cds, alternatively spliced
Length = 3670

Score = 867 bits (2215), Expect = 0.0
Identities = 425/425 (100%), Positives = 425/425 (100%)
Frame = +1

```
Query: 1    MTFGRSGAASVVLNVGGARYSLSRELLKDFPLRRVSRLHGCRSERDVLEVCDDYDRERNE 60
            MTFGRSGAASVVLNVGGARYSLSRELLKDFPLRRVSRLHGCRSERDVLEVCDDYDRERNE
Sbjct: 478 MTFGRSGAASVVLNVGGARYSLSRELLKDFPLRRVSRLHGCRSERDVLEVCDDYDRERNE 657

Query: 61   YFFDRHSEAFGFILLYVRGHGKLRFAPRMCELSFYNEMIYWGLEGAHLEYCCQRRLDDRM 120
            YFFDRHSEAFGFILLYVRGHGKLRFAPRMCELSFYNEMIYWGLEGAHLEYCCQRRLDDRM
Sbjct: 658 YFFDRHSEAFGFILLYVRGHGKLRFAPRMCELSFYNEMIYWGLEGAHLEYCCQRRLDDRM 837

Query: 121 SDTYTFYSADEPGVLGRDEARPGGAEAAPSRRWLERMRRTFEEPTSSLAAQILASVSVVF 180
            SDTYTFYSADEPGVLGRDEARPGGAEAAPSRRWLERMRRTFEEPTSSLAAQILASVSVVF
Sbjct: 838 SDTYTFYSADEPGVLGRDEARPGGAEAAPSRRWLERMRRTFEEPTSSLAAQILASVSVVF 1017

Query: 181 VIVSMVVLCASTLPDWRNAAADNRSLDDRSRIIEAICIGWFTAECIVRFIVSKNKCEFVK 240
            VIVSMVVLCASTLPDWRNAAADNRSLDDRSRIIEAICIGWFTAECIVRFIVSKNKCEFVK
Sbjct: 1018VIVSMVVLCASTLPDWRNAAADNRSLDDRSRIIEAICIGWFTAECIVRFIVSKNKCEFVK 1197

Query: 241 RPLNIIDLLAITPYYISVLMTVFTGENSQLQRAGVTLRVLRMMRIFWVIKLARHFIGLQT 300
            RPLNIIDLLAITPYYISVLMTVFTGENSQLQRAGVTLRVLRMMRIFWVIKLARHFIGLQT
Sbjct: 1198RPLNIIDLLAITPYYISVLMTVFTGENSQLQRAGVTLRVLRMMRIFWVIKLARHFIGLQT 1377

Query: 301 LGLTLKRCYREMVMLLVFICVAMAIFSALSQLLEHGLDLETSNKDFTSIPAACWWVIISM 360
            LGLTLKRCYREMVMLLVFICVAMAIFSALSQLLEHGLDLETSNKDFTSIPAACWWVIISM
Sbjct: 1378LGLTLKRCYREMVMLLVFICVAMAIFSALSQLLEHGLDLETSNKDFTSIPAACWWVIISM 1557

Query: 361 TTVGYGDMYPITVPGRILGGVCVVSGIVLLALPITFIYHSFVQCYHELKFRSARYSRSLS 420
            TTVGYGDMYPITVPGRILGGVCVVSGIVLLALPITFIYHSFVQCYHELKFRSARYSRSLS
Sbjct: 1558TTVGYGDMYPITVPGRILGGVCVVSGIVLLALPITFIYHSFVQCYHELKFRSARYSRSLS 1737

Query: 421 TEFLN 425
            TEFLN
Sbjct: 1738TEFLN 1752
```

☐ 1: AF454547. Reports Homo sapiens volt...[gi:22164081] Links

```
LOCUS       AF454547                3670 bp    mRNA    linear   PRI 09-JUL-2004
DEFINITION  Homo sapiens voltage-gated potassium channel subunit Kv10.1a mRNA,
            complete cds, alternatively spliced.
ACCESSION   AF454547
VERSION     AF454547.1  GI:22164081
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 3670)
  AUTHORS   Vega-Saenz de Miera,E.C.
  TITLE     Modification of Kv2.1 K+ currents by the silent Kv10 subunits
  JOURNAL   Brain Res. Mol. Brain Res. 123 (1-2), 91-103 (2004)
   PUBMED   15046870
REFERENCE   2  (bases 1 to 3670)
  AUTHORS   Vega-Saenz de Miera,E.C. and Rudy,B.
  TITLE     Kv10.1a and Kv10.1b: Two novel alternatively spliced potassium
            channel subunits
  JOURNAL   Unpublished
REFERENCE   3  (bases 1 to 3670)
  AUTHORS   Vega-Saenz de Miera,E.C. and Rudy,B.
  TITLE     Direct Submission
  JOURNAL   Submitted (04-DEC-2001) Physiology and Neuroscience, New York
            University School of Medicine, 550 First Avenue, New York, NY
            10016, USA
FEATURES             Location/Qualifiers
     source          1..3670
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /chromosome="2"
                     /map="2p22-p21"
     CDS             478..1755
                     /note="alternatively spliced"
                     /codon_start=1
                     /product="voltage-gated potassium channel subunit Kv10.1a"
                     /protein_id="AAM93548.1"
                     /db_xref="GI:22164082"
                     /translation="MTFGRSGAASVVLNVGGARYSLSRELLKDFPLRRVSRLHGCRSE
                     RDVLEVCDDYDRERNEYFFDRHSEAFGFILLYVRGHGKLRFAPRMCELSFYNEMIYWG
                     LEGAHLEYCCQRRLDDRMSDTYTFYSADEPGVLGRDEARPGGAEAAPSRRWLERMRRT
                     FEEPTSSLAAQILASVSVVFVIVSMVVLCASTLPDWRNAAADNRSLDDRSRIIEAICI
                     GWFTAECIVRFIVSKNKCEFVKRPLNIIDLLAITPYYISVLMTVFTGENSQLQRAGVT
```

```
                     LRVLRMMRIFWVIKLARHFIGLQTLGLTLKRCYREMVMLLVFICVAMAIFSALSQLLE
                     HGLDLETSNKDFTSIPAACWWVIISMTTVGYGDMYPITVPGRILGGVCVVSGIVLLAL
                     PITFIYHSFVQCYHELKFRSARYSRSLSTEFLN"
     polyA_signal    3650..3655
     polyA_site      3670
ORIGIN
        1 ggcctctcgc ctctggacgg cggcggggcg gccgccggat tcgcggccgc agggagcgcc
       61 ggagacgggg agctattccg ccccggcggc tccattcggc gcccgcagcc ctcagggggt
      121 cggccccgcg gcttgggaga gggcaccgcg gcctcggtgt gcgcagccct cgggcgcgag
      181 ggtcggcggc gcggacacag ccgcgttccc agccggtggg gctcagcgct ggcgccggca
      241 aggactcccc ggccacccgc aggtaccgcc gggcggaggg cgcgctacta gcagcgccgg
      301 agatactcga gcccagggac ccccgggcca gcggagggca ggagcggagc cccgagggag
      361 cgcgggcccc gacggcgcgc tcccccgtca gccacgggca ggcaggcccc gcgtggcggc
      421 ttggggtggg gggctgcagc ggggccctcg ggccgaaagt cccccgggcg gccagccatg
      481 accttcgggc gcagcggggc ggcctcggtg gtgctgaacg tgggcggcgc ccggtattcg
      541 ctgtcccggg agctgctgaa ggacttcccg ctgcgccgcg tgagccggct gcacggctgc
      601 cgctccgagc gcgacgtgct cgaggtgtgc gacgactacg accgcgagcg caacgagtac
      661 ttcttcgacc ggcactcgga ggccttcggc ttcatcctgc tctacgtgcg cggccacggc
      721 aagctgcgct tcgcgccgcg gatgtgcgag ctctccttct acaacgagat gatctactgg
      781 ggcctggagg gcgcgcacct cgagtactgc tgccagcgcc gcctcgacga ccgcatgtcc
      841 gacacctaca ccttctactc ggccgacgag ccgggcgtgc tgggccgcga cgaggcgcgc
      901 cccggcgggg ccgaggcggc tccctccagg cgctggctgg agcgcatgcg gcggaccttc
      961 gaggagccca cgtcgtcgct ggccgcgcag atcctggcta gcgtgtcggt ggtgttcgtg
     1021 atcgtgtcca tggtggtgct gtgcgccagc acgttgcccg actggcgcaa cgcagccgcc
     1081 gacaaccgca gcctggatga ccggagcagg ataattgaag ctatctgcat aggttggttc
     1141 actgccgagt gcatcgtgag gttcattgtc tccaaaaaca agtgtgagtt tgtcaagaga
     1201 cccctgaaca tcattgattt actggcaatc acgccgtatt acatctctgt gttgatgaca
     1261 gtgtttacag gcgagaactc tcaactccag agggctggag tcaccttgag ggtacttaga
     1321 atgatgagga ttttttgggt gattaagctt gcccgtcact tcattggtct tcagacactc
     1381 ggtttgactc tcaaacgttg ctaccgagag atggttatgt tacttgtctt catttgtgtt
     1441 gccatggcaa tctttagtgc actttctcag cttcttgaac atgggctgga cctggaaaca
     1501 tccaacaagg actttaccag cattcctgct gcctgctggt gggtgattat ctctatgact
     1561 acagttggct atggagatat gtatcctatc acagtgcctg gaagaattct tggaggagtt
     1621 tgtgttgtca gtggaattgt tctattggca ttacctatca cttttatcta ccatagcttt
     1681 gtgcagtgtt atcatgagct caagtttaga tctgctaggt atagtaggag cctctccact
     1741 gaattcctga attaatgcat tgcaaatcaa ttcttgcata cacttcatag aaagactttg
     1801 atgctgcttc atatttatgt gtttcttgct gggtgagcac tgcagtggca ttgtcatcat
     1861 cttggtaggg taaaaattat ccttcccagc cgaagggata aaacagttta cttgttatgg
     1921 agtaaataga attgagactg caaaggaaga ataatgactc ctagagtaaa ctttaggacc
     1981 cggttttatt tagacttgtt ttcccgtttc cttgaatgat tacacatttt taaaaaatac
     2041 attatttgaa cattttaaaa cagaaaggta ctattttcca aatgtttttc catcttatga
     2101 attcagtaga agcttggaac ttatagtgtt ttttgtttga gagtaacatt ttcatttcta
     2161 aatgttttat aatttctcat atcaatgtca gaagtatcct ggaaacatat gtcacatgcg
     2221 ggaactgttt aacaaatact ttaaaaattt ggccaaaatt taaactgtat aatggagcta
     2281 gatacaagca agaatagtat ttgaaagact tttccagcat acttctcaat tctttgcttt
     2341 atttttgtgc caattattca ccttatcgtg ccgcttcatg gaagcttgag tatgttctcc
     2401 cttttccatt ttggatttat ctctttactg taatgactca aaaggtattt aagaattgac
     2461 gagagcttgt gttgtttagc atcttactgg ataatatttg aattcattgc tgttcctagg
     2521 tgataactgt cctaatattt agatgtccaa acaagaatac ttccaacata aaaattataa
     2581 taggaataat ttgagatgac tcaatattac aacctcttct tctcttaacc tcctcccca
     2641 aacactagag gtttaataag acttatcaga tgaaaggata tttatatagc cttttagtag
     2701 caaagtcata cttacgtgtt gtcactggat tatcataaaa gggagaaatt aaatattact
     2761 gtactcttag ttgctgtgta gctaagtcaa ttttaagcca gtaaaagcga tggatacata
     2821 atgatttgat ctgatcttta actattgtga atcacagcta caccaaaact cttcttgtaa
     2881 gaatactgac taatatgcca tgttaatctg gctagattat taggactaga taatgtaaaa
     2941 gtgatgattg tttagtaact aaattttagc aacagaaatt agaattttgc tttttcaacc
     3001 agttaccata aagaagttag tgtatatata aacacaaata attagtgaca gattcataaa
     3061 aaattgaatg ttgtacacag taattttgtc agaggtagag aagacaggga ttgggaagtg
     3121 gtgggtgatg gaggacctgg atatatttat caaataaagg gttaccagaa gtgttcatta
     3181 aaggaatttt agccatcatc tagttcaagc ctcaactatt acaggtagaa aatcagggca
```

```
3241 ggagagaata taattgtgaa ggagtcaggg ctaacacctg gatctccaga aacctagccc
3301 agcaggttaa tcttcacaca tctctgggtt ctgagaaaag cctggaaaaa tcacacttct
3361 ttgtcattgt catgctgagg taataatagc aaaactgttt tctttccctt aatttccttt
3421 cctaagctta tgtaatagtt tggccattaa atatcttgcc ctattttccc tattactgct
3481 agtatgctac ttcttacata cccaaaagaa attcagttat ttattgtata tttattgtat
3541 tctaatataa ttgaaataaa tggcatggat ttatttttc ttaactattt ggattaaagc
3601 tttgtggttc atgcaaacaa tgtgcagatg atagcacctc catattacta ataaaaatat
3661 gataaccatc
//
```


Feb 9 2005 14:31:10

EXHIBIT C

>NM_172344 ACCESSION:NM_172344 NID: gi 27436992 ref NM_172344.1 Homo sapiens potassium voltage-gated channel, subfamily G, member 3 (KCNG3), transcript variant 2, mRNA
Length = 3791

Score = 867 bits (2215), Expect = 0.0
Identities = 425/425 (100%), Positives = 425/425 (100%)
Frame = +3

```
Query: 1    MTFGRSGAASVVLNVGGARYSLSRELLKDFPLRRVSRLHGCRSERDVLEVCDDYDRERNE 60
            MTFGRSGAASVVLNVGGARYSLSRELLKDFPLRRVSRLHGCRSERDVLEVCDDYDRERNE
Sbjct: 597  MTFGRSGAASVVLNVGGARYSLSRELLKDFPLRRVSRLHGCRSERDVLEVCDDYDRERNE 776

Query: 61   YFFDRHSEAFGFILLYVRGHGKLRFAPRMCELSFYNEMIYWGLEGAHLEYCCQRRLDDRM 120
            YFFDRHSEAFGFILLYVRGHGKLRFAPRMCELSFYNEMIYWGLEGAHLEYCCQRRLDDRM
Sbjct: 777  YFFDRHSEAFGFILLYVRGHGKLRFAPRMCELSFYNEMIYWGLEGAHLEYCCQRRLDDRM 956

Query: 121  SDTYTFYSADEPGVLGRDEARPGGAEAAPSRRWLERMRRTFEEPTSSLAAQILASVSVVF 180
            SDTYTFYSADEPGVLGRDEARPGGAEAAPSRRWLERMRRTFEEPTSSLAAQILASVSVVF
Sbjct: 957  SDTYTFYSADEPGVLGRDEARPGGAEAAPSRRWLERMRRTFEEPTSSLAAQILASVSVVF 1136

Query: 181  VIVSMVVLCASTLPDWRNAAADNRSLDDRSRIIEAICIGWFTAECIVRFIVSKNKCEFVK 240
            VIVSMVVLCASTLPDWRNAAADNRSLDDRSRIIEAICIGWFTAECIVRFIVSKNKCEFVK
Sbjct: 1137 VIVSMVVLCASTLPDWRNAAADNRSLDDRSRIIEAICIGWFTAECIVRFIVSKNKCEFVK 1316

Query: 241  RPLNIIDLLAITPYYISVLMTVFTGENSQLQRAGVTLRVLRMMRIFWVIKLARHFIGLQT 300
            RPLNIIDLLAITPYYISVLMTVFTGENSQLQRAGVTLRVLRMMRIFWVIKLARHFIGLQT
Sbjct: 1317 RPLNIIDLLAITPYYISVLMTVFTGENSQLQRAGVTLRVLRMMRIFWVIKLARHFIGLQT 1496

Query: 301  LGLTLKRCYREMVMLLVFICVAMAIFSALSQLLEHGLDLETSNKDFTSIPAACWWVIISM 360
            LGLTLKRCYREMVMLLVFICVAMAIFSALSQLLEHGLDLETSNKDFTSIPAACWWVIISM
Sbjct: 1497 LGLTLKRCYREMVMLLVFICVAMAIFSALSQLLEHGLDLETSNKDFTSIPAACWWVIISM 1676

Query: 361  TTVGYGDMYPITVPGRILGGVCVVSGIVLLALPITFIYHSFVQCYHELKFRSARYSRSLS 420
            TTVGYGDMYPITVPGRILGGVCVVSGIVLLALPITFIYHSFVQCYHELKFRSARYSRSLS
Sbjct: 1677 TTVGYGDMYPITVPGRILGGVCVVSGIVLLALPITFIYHSFVQCYHELKFRSARYSRSLS 1856

Query: 421  TEFLN 425
            TEFLN
Sbjct: 1857 TEFLN 1871
```

☐ 1: NM_172344. Reports Homo sapiens pota...[gi:27436992] Links

LOCUS NM_172344 3791 bp mRNA linear PRI 20-DEC-2004
DEFINITION Homo sapiens potassium voltage-gated channel, subfamily G, member 3 (KCNG3), transcript variant 2, mRNA.
ACCESSION NM_172344
VERSION NM_172344.1 GI:27436992
KEYWORDS .
SOURCE Homo sapiens (human)
ORGANISM Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE 1 (bases 1 to 3791)
AUTHORS Vega-Saenz de Miera,E.C.
TITLE Modification of Kv2.1 K+ currents by the silent Kv10 subunits
JOURNAL Brain Res. Mol. Brain Res. 123 (1-2), 91-103 (2004)
PUBMED 15046870
REFERENCE 2 (bases 1 to 3791)
AUTHORS Ottschytsch,N., Raes,A., Van Hoorick,D. and Snyders,D.J.
TITLE Obligatory heterotetramerization of three previously uncharacterized Kv channel alpha-subunits identified in the human genome
JOURNAL Proc. Natl. Acad. Sci. U.S.A. 99 (12), 7986-7991 (2002)
MEDLINE 22056098
PUBMED 12060745
REMARK GeneRIF: Obligatory heterotetramerization of three previously uncharacterized Kv channel subunits identified in human genome (Kv6.3)(Kv10.1) (Kv11.1)
GeneRIF: Obligatory heterotetramerization of three previously uncharacterized Kv channel subunits identified in human genome
REFERENCE 3 (bases 1 to 3791)
AUTHORS Sano,Y., Mochizuki,S., Miyake,A., Kitada,C., Inamura,K., Yokoi,H., Nozawa,K., Matsushime,H. and Furuichi,K.
TITLE Molecular cloning and characterization of Kv6.3, a novel modulatory subunit for voltage-gated K(+) channel Kv2.1
JOURNAL FEBS Lett. 512 (1-3), 230-234 (2002)
MEDLINE 21841130
PUBMED 11852086
REMARK GeneRIF: These results indicate that Kv6.3 is a novel member of the voltage-gated K(+) channel which functions as a modulatory subunit of the Kv2.1 channel.
COMMENT REVIEWED REFSEQ: This record has been curated by NCBI staff. The reference sequence was derived from AF454547.1 and AF348982.1.

Summary: Voltage-gated potassium (Kv) channels represent the most complex class of voltage-gated ion channels from both functional

and structural standpoints. Their diverse functions include regulating neurotransmitter release, heart rate, insulin secretion, neuronal excitability, epithelial electrolyte transport, smooth muscle contraction, and cell volume. This gene encodes a member of the potassium channel, voltage-gated, subfamily G. This member is a gamma subunit functioning as a modulatory molecule. Alternative splicing results in two transcript variants encoding distinct isoforms.

Transcript Variant: This variant (2), also known as Kv10.1a, lacks an alternate in-frame segment in the coding region, as compared to variant 1. It encodes isoform (2) that lacks an internal segment, as compared to isoform 1.
COMPLETENESS: complete on the 3' end.

FEATURES Location/Qualifiers

source 1..3791
/organism="Homo sapiens"
/mol_type="mRNA"
/db_xref="taxon:9606"
/chromosome="2"
/map="2p21"

gene 1..3791
/gene="KCNG3"
/note="synonyms: KV6.3, KV10.1"
/db_xref="GeneID:170850"
/db_xref="MIM:606767"

CDS 597..1874
/gene="KCNG3"
/note="isoform 2 is encoded by transcript variant 2; voltage-gated potassium channel 6.3; voltage-gated potassium channel Kv10.1; voltage-gated potassium channel subunit Kv6.4;
go_component: plasma membrane [goid 0005886] [evidence IDA] [pmid 12060745];
go_component: integral to membrane [goid 0016021] [evidence IEA];
go_component: endoplasmic reticulum [goid 0005783] [evidence IDA] [pmid 12060745];
go_component: voltage-gated potassium channel complex [goid 0008076] [evidence IEA];
go_function: protein binding [goid 0005515] [evidence IPI] [pmid 11852086];
go_function: voltage-gated potassium channel activity [goid 0005249] [evidence IEA];
go_process: cation transport [goid 0006812] [evidence IEA];
go_process: potassium ion transport [goid 0006813] [evidence IEA]"
/codon_start=1
/product="potassium voltage-gated channel, subfamily G, member 3 isoform 2"
/protein_id="NP_758847.1"
/db_xref="GI:27436993"
/db_xref="GeneID:170850"
/db_xref="MIM:606767"
/translation="MTFGRSGAASVVLNVGGARYSLSRELLKDFPLRRVSRLHGCRSE
RDVLEVCDDYDRERNEYFFDRHSEAFGFILLYVRGHGKLRFAPRMCELSFYNEMIYWG
LEGAHLEYCCQRRLDDRMSDTYTFYSADEPGVLGRDEARPGGAEAAPSRRWLERMRRT
FEEPTSSLAAQILASVSVVFVIVSMVVLCASTLPDWRNAAADNRSLDDRSRIIEAICI
GWFTAECIVRFIVSKNKCEFVKRPLNIIDLLAITPYYISVLMTVFTGENSQLQRAGVT

```
                     LRVLRMMRIFWVIKLARHFIGLQTLGLTLKRCYREMVMLLVFICVAMAIFSALSQLLE
                     HGLDLETSNKDFTSIPAACWWVIISMTTVGYGDMYPITVPGRILGGVCVVSGIVLLAL
                     PITFIYHSFVQCYHELKFRSARYSRSLSTEFLN"
     polyA_signal    3769..3774
                     /gene="KCNG3"
     polyA_site      3791
                     /gene="KCNG3"
ORIGIN
        1 gcggcgcgga gggaggtgag cggcgcgcgc ggagccggcg ggcgaggagg aggactgcac
       61 agaggccccg cccccgccgc cgcgagccgg ctcttcgccg cctccgaacc cgctcacttt
      121 gcctctcgcc tctggacggc ggcggggcgg ccgccggatt cgcggccgca gggagcgccg
      181 gagacgggga gctattccgc cccggcggct ccattcggcg cccgcagccc tcaggggtc
      241 ggccccgcgg cttgggagag ggcaccgcgg cctcggtgtg cgcagccctc gggcgcgagg
      301 gtcggcggcg cggacacagc cgcgttccca gccggtgggg ctcagcgctg gcgccggcga
      361 ggactccccg gccacccgca ggtaccgccg ggcggagggc gcgctactag cagcgccgga
      421 gatactcgag cccagggacc cccgggccag cggagggcag gagcggagcc ccgagggagc
      481 gcgggccccg acggcgcgct cccccgtcag ccacgggcag gcaggccccg cgtggcggct
      541 tggggtgggg ggctgcagcg gggccctcgg gccgaaagtc ccccgggcgg ccagccatga
      601 ccttcgggcg cagcggggcg gcctcggtgg tgctgaacgt gggcggcgcc cggtattcgc
      661 tgtcccggga gctgctgaag gacttcccgc tgcgccgcgt gagccggctg cacggctgcc
      721 gctccgagcg cgacgtgctc gaggtgtgcg acgactacga ccgcgagcgc aacgagtact
      781 tcttcgaccg gcactcggag gccttcggct tcatcctgct ctacgtgcgc ggccacggca
      841 agctgcgctt cgcgccgcgg atgtgcgagc tctccttcta caacgagatg atctactggg
      901 gcctggaggg cgcgcacctc gagtactgct gccagcgccg cctcgacgac cgcatgtccg
      961 acacctacac cttctactcg gccgacgagc cgggcgtgct gggccgcgac gaggcgcgcc
     1021 ccggcggggc cgaggcggct ccctccaggc gctggctgga gcgcatgcgg cggaccttcg
     1081 aggagcccac gtcgtcgctg gccgcgcaga tcctggctag cgtgtcggtg gtgttcgtga
     1141 tcgtgtccat ggtggtgctg tgcgccagca cgttgcccga ctggcgcaac gcagccgccg
     1201 acaaccgcag cctggatgac cggagcagga taattgaagc tatctgcata ggttggttca
     1261 ctgccgagtg catcgtgagg ttcattgtct ccaaaaacaa gtgtgagttt gtcaagagac
     1321 ccctgaacat cattgattta ctggcaatca cgccgtatta catctctgtg ttgatgacag
     1381 tgtttacagg cgagaactct caactccaga gggctggagt caccttgagg gtacttagaa
     1441 tgatgaggat tttttgggtg attaagcttg cccgtcactt cattggtctt cagacactcg
     1501 gtttgactct caaacgttgc taccgagaga tggttatgtt acttgtcttc atttgtgttg
     1561 ccatggcaat ctttagtgca ctttctcagc ttcttgaaca tgggctggac ctggaaacat
     1621 ccaacaagga ctttaccagc attcctgctg cctgctggtg ggtgattatc tctatgacta
     1681 cagttggcta tggagatatg tatcctatca cagtgcctgg aagaattctt ggaggagttt
     1741 gtgttgtcag tggaattgtt ctattggcat tacctatcac ttttatctac catagctttg
     1801 tgcagtgtta tcatgagctc aagtttagat ctgctaggta tagtaggagc ctctccactg
     1861 aattcctgaa ttaatgcatt gcaaatcaat tcttgcatac acttcataga aagactttga
     1921 tgctgcttca tatttatgtg tttcttgctg ggtgagcact gcagtggcat tgtcatcatc
     1981 ttggtagggt aaaaattatc cttcccagcc gaagggataa aacagtttac ttgttatgga
     2041 gtaaatagaa ttgagactgc aaaggaagaa taatgactcc tagagtaaac tttaggaccc
     2101 ggttttattt agacttgttt tcccgtttcc ttgaatgatt acacattttt aaaaaataca
     2161 ttatttgaac attttaaaac agaaaggtac tattttccaa atgtttttcc atcttatgaa
     2221 ttcagaagaa gcttggaact tatagtgttt tttgtttgag agtaacattt tcatttctaa
     2281 atgttttata atttctcata tcaatgtcag aagtatcctg gaaacatatg tcacatgcgg
     2341 gaactgttta acaaatactt taaaaatttg gccaaaattt aaactgtata atggagctag
     2401 atacaagcaa gaatagtatt tgaaagactt ttccagcata cttctcaatt ctttgcttta
     2461 tttttgtgcc aattattcac cttatcgtgc cgcttcatgg aagcttgagt atgttctccc
     2521 ttttccattt tggatttatc tctttactgt aatgactcaa aaggtattta agaattgacg
     2581 agagcttgtg ttgtttagca tcttactgga taatatttga attcattgct gttcctaggt
     2641 gataactgtc ctaatattta gatgtccaaa caagaatact tccaacataa aaattataat
     2701 aggaataatt tgagatgact caatattaca acctcttctt ctcttaacct cctcccccaa
     2761 acactagagg tttaataaga cttatcagat gaaaggatat ttatatagcc ttttagtagc
     2821 aaagtcatac ttacgtgttg tcactggatt atcataaaag ggagaaatta aatattactg
     2881 tactcttagt tgctgtgtag ctaagtcaat tttaagccag taaaagcgat ggatacataa
     2941 tgatttgatc tgatctttaa ctattgtgaa tcacagctac accaaaactc ttcttgtaag
     3001 aatactgact aatatgccat gttaatctgg ctagattatt aggactagat aatgtaaaag
     3061 tgatgattgt ttagtaacta aattttagca acagaaatta gaattttgct ttttcaacca
```

```
3121 gttaccataa agaagttagt gtatatataa acacaaataa ttagtgacag attcataaaa
3181 aattgaatgt tgtacacagt aattttgtca gaggtagaga agacagggat tgggaagtgg
3241 tgggtgatgg aggacctgga tatatttatc aaataaaggg ttaccagaag tgttcattaa
3301 aggaatttta gccatcatct agttcaaacc tcaactatta caggtagaaa atcagggcag
3361 gagagaatat aattgtgaag gagtcagggc taacacctgg atctccagaa acctagccca
3421 gcaggttaat cttcacacat ctctgggttc tgagaaaagc ctggaaaaat cacacttctt
3481 tgtcattgtc atgctgaggt aataatagca aaactgtttt ctttcccttа atttcctttc
3541 ctaagcttat gtaatagttt ggccattaaa tatcttgccc tattttccct attactgcta
3601 gtatgctact tcttacatac ccaaaagaaa ttcagttatt tattgtatat ttattgtatt
3661 ctaatataat tgaaataaat ggcatggatt tatttttct taactatttg gattaaagct
3721 ttgtggttca tgcaaacaat gtgcagatga tagcacctcc atattactaa taaaaatatg
3781 ataaccatca a
//
```


Feb 9 2005 14:31:10

EXHIBIT D

1: FEBS Lett. 2002 Feb 13;512(1-3):230-4.

Molecular cloning and characterization of Kv6.3, a novel modulatory subunit for voltage-gated K(+) channel Kv2.1.

Sano Y, Mochizuki S, Miyake A, Kitada C, Inamura K, Yokoi H, Nozawa K, Matsushime H, Furuichi K.

Molecular Medicine Laboratories, Institute for Drug Discovery Research, Yamanouchi Pharmaceutical Co., Ltd., 21 Miyukigaoka, Ibaraki 305-8585, Tsukuba, Japan. sano.yorikata@yamanouchi.co.jp

We report identification and characterization of Kv6.3, a novel member of the voltage-gated K(+) channel. Reverse transcriptase-polymerase chain reaction analysis indicated that Kv6.3 was highly expressed in the brain. Electrophysiological studies indicated that homomultimeric Kv6.3 did not yield a functional voltage-gated ion channel. When Kv6.3 and Kv2.1 were co-expressed, the heteromultimeric channels displayed the decreased rate of deactivation compared to the homomultimeric Kv2.1 channels. Immunoprecipitation studies indicated that Kv6.3 bound with Kv2.1 in co-transfected cells. These results indicate that Kv6.3 is a novel member of the voltage-gated K(+) channel which functions as a modulatory subunit.

MeSH Terms:

- Amino Acid Sequence
- Cloning, Molecular
- Electric Conductivity
- Humans
- Ion Channel Gating
- Molecular Sequence Data
- Potassium Channels/classification
- Potassium Channels/genetics
- Potassium Channels/metabolism*
- Potassium Channels, Voltage-Gated*
- Protein Subunits
- Sequence Homology, Amino Acid
- Tissue Distribution

Substances:

- Potassium Channels
- Potassium Channels, Voltage-Gated
- Protein Subunits
- delayed rectifier potassium channel
- potassium channel Kv6.3

Secondary Source ID:

- GENBANK/AB070604
- GENBANK/AB070605

PMID: 11852086 [PubMed - indexed for MEDLINE]

Obligatory heterotetramerization of three previously uncharacterized Kv channel α-subunits identified in the human genome

N. Ottschytsch, A. Raes, D. Van Hoorick, and D. J. Snyders*

Laboratory for Molecular Biophysics, Physiology, and Pharmacology, University of Antwerp (UIA) and Flanders Institute for Biotechnology (VIB), B2610 Antwerp, Belgium

Edited by Lily Y. Jan, University of California School of Medicine, San Francisco, CA, and approved April 12, 2002 (received for review November 20, 2001)

Voltage-gated K^+ channels control excitability in neuronal and various other tissues. We identified three unique α-subunits of voltage-gated K^+-channels in the human genome. Analysis of the full-length sequences indicated that one represents a previously uncharacterized member of the Kv6 subfamily, Kv6.3, whereas the others are the first members of two unique subfamilies, Kv10.1 and Kv11.1. Although they have all of the hallmarks of voltage-gated K^+ channel subunits, they did not produce K^+ currents when expressed in mammalian cells. Confocal microscopy showed that Kv6.3, Kv10.1, and Kv11.1 alone did not reach the plasma membrane, but were retained in the endoplasmic reticulum. Yeast two-hybrid experiments failed to show homotetrameric interactions, but showed interactions with Kv2.1, Kv3.1, and Kv5.1. Co-expression of each of the previously uncharacterized subunits with Kv2.1 resulted in plasma membrane localization with currents that differed from typical Kv2.1 currents. This heteromerization was confirmed by co-immunoprecipitation. The Kv2 subfamily consists of only two members and uses interaction with "silent subunits" to diversify its function. Including the subunits described here, the "silent subunits" represent one-third of all Kv subunits, suggesting that obligatory heterotetramer formation is more widespread than previously thought.

electrically silent subunits | ER retention | heterotetrameric assembly | KCNG3

Voltage-gated potassium channels are transmembrane proteins consisting of four α-subunits that form a central permeation pathway. Each subunit contains six transmembrane domains (S1–S6) and a pore loop containing the GYG-motif, the signature sequence for potassium selectivity. The fourth transmembrane domain (S4) contains positively charged residues and is the major part of the voltage sensor. Voltage-gated potassium channels serve a wide range of functions including regulation of the resting membrane potential and control of the shape, duration, and frequency of action potentials (1–3).

At present, 26 genes have been described encoding for different Kv α-subunits. These are divided into subfamilies by sequence similarities: within a subfamily members share ≈70% of sequence identity, whereas between different subfamilies this percentage drops to ≈40%, reflecting the homology in the core section S1–S6 (4). The Kv family of potassium channels consists of nine subfamilies, Kv1 through Kv9, although Kv7 has only been described for Aplysia (5). The subunits of the Kv1 through Kv4 subfamilies all show functional expression in a homotetrameric configuration. Despite having the typical topology of voltage-gated potassium channel subunits, the subunits of the Kv5 through Kv9 families cannot generate current by themselves (6–10). For instance, Kv6.1 fails to form homotetrameric channels, but it is able to form heterotetrameric channels with Kv2.1; expression of these heterotetramers resulted in currents with clearly distinguishable properties (11). All known "electrically silent" subunits have been shown to form heterotetrameric channels with the members of the Kv2 subfamily (8–10). In a sense, these "silent" subunits can be considered regulatory subunits—e.g., the metabolic regulation of the Kv2.1/Kv9.3 heteromultimer might play an important role in hypoxic pulmonary artery vasoconstriction and in the possible development of pulmonary hypertension (8).

In this study we report the cloning and functional properties of three previously uncharacterized subunits that were identified in the early public draft version of the human genome. Based on sequence identity, one of these is a previously uncharacterized member of the Kv6 subfamily (Kv6.3), whereas the others are the first members of two unique subfamilies, Kv10.1 and Kv11.1. Biochemical, microscopic, and functional analysis indicated that these previously uncharacterized subunits are all "silent subunits," which may explain why they have not been cloned previously. Through obligatory heterotetramerization they exert a function-altering effect on other Kv subunits.

Experimental Procedures

Cloning of Kv2.1, Kv6.3, Kv10.1, and Kv11.1. The coding sequence of human Kv2.1 was amplified from a human brain library (CLONTECH) and cloned into pEGFP-N1. The channel sequences of Kv6.3, Kv10.1, and Kv11.1 were obtained through a BLAST search of the high throughput genomic sequence (*htgs*) database (July 2000). The coding sequences were cloned using PCR amplification from a human brain library (CLONTECH) or a human testis library (TaKaRa, Shiga, Japan) for Kv10.1 and Kv11.1, respectively. Both coding exons of Kv6.3 were amplified from human genomic DNA. The BsaMI restriction site at the start of the second exon was used to join the two coding exons.

Amino Acid Sequence Alignments and Phylogenetic Tree. Computer analyses were performed using MEGALIGN (DNAstar, Madison, WI). The phylogenetic tree and the percentage of identity were obtained by aligning the core S1–S6 sequences (e.g., aa 252–518 in Kv1.5).

Expression Analysis. A cDNA panel from different tissues was obtained from CLONTECH (cDNA panel I and II). PCR was performed with primer sets that were selected to ensure the amplification of the correct subunit, without amplification of homologous subunits. All reactions were done with 38 cycles and PCR products were analyzed on a 2% agarose gel.

Transfection. Ltk^- cells were cultured and transfected with cDNA as reported (12). Each subunit was coexpressed with Kv2.1 (10:1

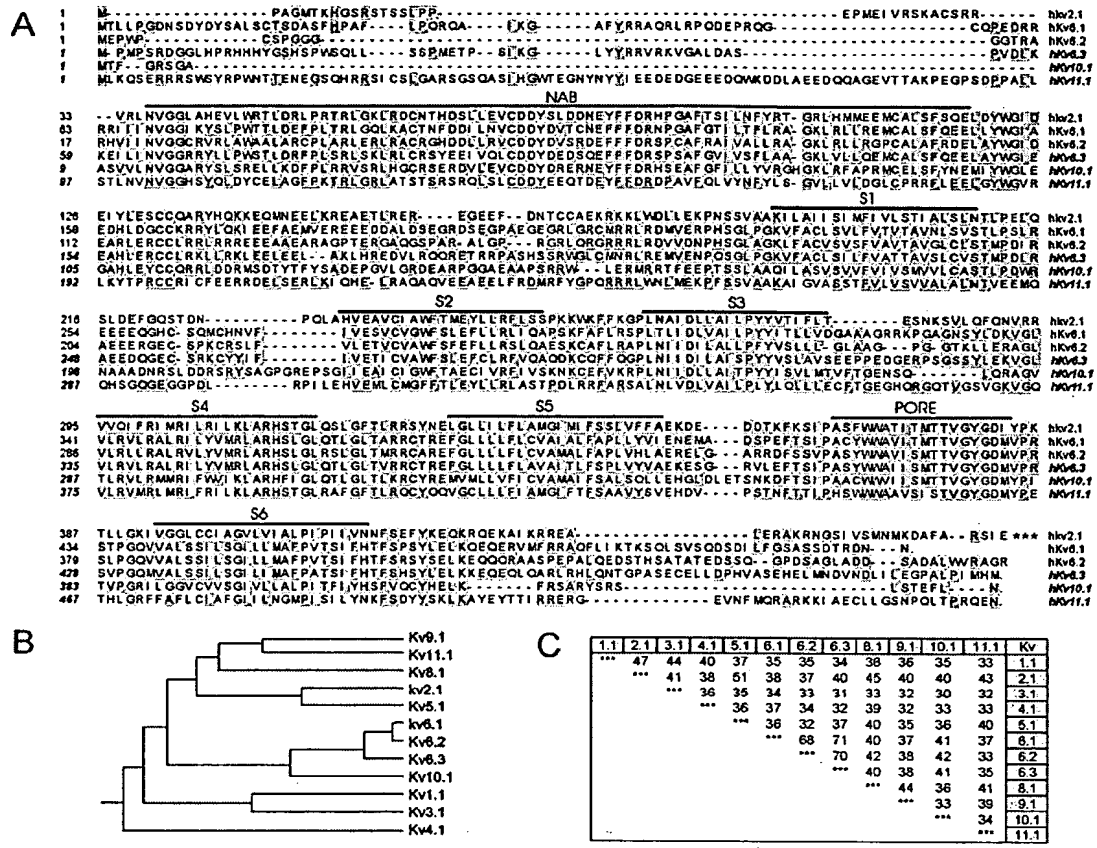
This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: ER, endoplasmic reticulum; GFP, green fluorescent protein.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. AF348982–AF348984).

*To whom reprint requests should be addressed at: Laboratory for Molecular Biophysics, Physiology, and Pharmacology, Department of Biomedical Sciences, University of Antwerp (UIA), Universiteitsplein 1, T4.21, 2610 Antwerp, Belgium. E-mail: dirk.snyders@ua.ac.be.

www.pnas.org/cgi/doi/10.1073/pnas.122617999



	1.1	2.1	3.1	4.1	5.1	6.1	6.2	6.3	8.1	9.1	10.1	11.1	Kv
1.1	***	47	44	40	37	35	35	34	38	36	35	33	1.1
2.1		***	41	38	51	38	37	40	45	40	40	43	2.1
3.1			***	36	35	34	33	31	33	32	30	32	3.1
4.1				***	36	37	34	32	39	32	33	33	4.1
5.1					***	36	32	37	40	35	36	40	5.1
6.1						***	68	71	40	37	41	37	6.1
6.2							***	70	42	38	42	33	6.2
6.3								***	40	38	41	35	6.3
8.1									***	44	36	41	8.1
9.1										***	33	39	9.1
10.1											***	34	10.1
11.1												***	11.1

Fig. 1. Sequence alignment, phylogenetic tree, and percent sequence identity of Kv6.3, Kv10.1, and Kv11.1. (*A*) The amino acid sequences of Kv2.1, Kv6.1, Kv6.2, Kv6.3, Kv10.1, and Kv11.1 were aligned using MEGALIGN. For convenience, only the first 460 aa of Kv2.1 are shown. Gaps (indicated by dashes) were introduced in the sequence to maintain the alignment. Conserved amino acids are shaded in gray. The six putative transmembrane domains and the pore region are indicated by an overline. (*B*) The phylogenetic tree for the Kv family. (*C*) The percent sequence similarity based on the S1–S6 core.

ratio). At this ratio, less than 0.01% of the channels will be wild-type Kv2.1. Between 12 and 24 h post-transfection the cells were trypsinized and used for analysis.

Whole-Cell Current Recording. Current recordings were made with an Axopatch-200B amplifier (Axon instruments, Union City, CA) in the whole cell configuration of the patch-clamp technique (13) as reported (12).

Pulse Protocols and Data Analysis. The applied pulse protocols are listed in the figure legends. The voltage dependence of channel opening and inactivation (activation and inactivation curves) was fitted with a Boltzmann equation according to $y = 1/\{1 + exp[-(E - V_{1/2})/k]\}$, where $V_{1/2}$ represents the voltage at which 50% of the channels are open or inactivated and k the slope factor. Activation and deactivation kinetics were fitted with a single or double exponential function by using a nonlinear least-squares (Gauss-Newton) algorithm. Results are presented as mean ± SEM; statistical analysis was done using the Student's t test; probability values are presented in the text.

Yeast Two-Hybrid System and Protein Constructs. The MATCHMAKER Yeast Two-Hybrid System 3 (CLONTECH) was used to assay for protein–protein interactions. The amino termini of Kv1.5, Kv2.1, Kv3.1, Kv4.3, Kv5.1, Kv6.1, Kv6.3, Kv8.1, Kv9.3, Kv10.1, and Kv11.1 were cloned into the vector pGBKT7. The amino termini of Kv2.1, Kv6.3, Kv10.1, and Kv11.1 were also cloned into the vector pGADT7. AH109 cells were transformed with the plasmid constructs of interest (100 ng of each) and plated on −Trp/−Leu/+XαGAL media to select for cells containing both vectors and to test for interaction. The degree of interaction was determined from the speed and intensity of the blue color development.

Coimmunoprecipitation. Kv2.1 was c-myc-tagged at the C terminus and cotransfected with green fluorescent protein (GFP)-tagged Kv2.1, Kv1.5, Kv6.3, Kv10.1, or Kv11.1 into HEK293 cells. The next day the cells were solubilized on ice with a PBS buffer supplemented with 5 mM EDTA, 1% Triton X-100, and a complete protease inhibitor mixture (Roche Diagnostics). For the immunoprecipitation Protein G Agarose beads and 2 μg of anti-GFP (CLONTECH) were added. The samples were incubated overnight at 4°C with rocking. Beads were then washed with ice-cold solubilization buffer. Proteins were eluted from the beads by boiling in SDS sample buffer and analyzed on 8% SDS/PAGE. Proteins were transferred to a polyvinylidene difluoride (PVDF) membrane (Amersham Pharmacia Biotech) and the blot was blocked. The blot was incubated with anti-c-myc (CLONTECH); afterwards, anti-mouse IgG (Amersham Pharmacia Biotech) was added, followed by ECL detection (Amersham Pharmacia Biotech).

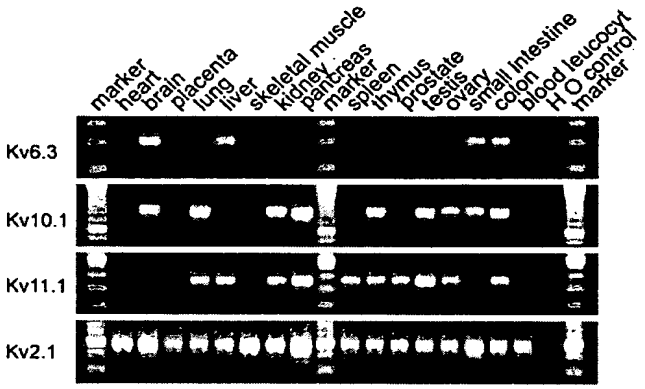


Fig. 2. Expression of Kv6.3, Kv10.1, Kv11.1, and Kv2.1 in human tissues. A PCR analysis was performed on a cDNA panel of the indicated human tissues with gene-specific primers for the subunits indicated on the left.

Confocal Imaging. Kv6.3, Kv10.1, and Kv11.1 were tagged with GFP at their carboxy terminus. HEK293 cells were cultivated on coverslips. For cotransfections, a 1:10 ratio of channel DNA versus Kv2.1 was added. The endoplasmic reticulum (ER) was visualized with the DsRed ER localization vector. This was constructed starting from the pDsRed vector (CLONTECH). The first 17 aa from calreticulin were amplified from brain cDNA and cloned in frame with the DsRed sequence of pDsRed. The KDEL sequence was inserted behind DsRed by using a mutagenesis PCR. Transfections and cotransfections (ratio 1:10 GFP-labeled channel DNA versus unlabeled Kv2.1 DNA) were done using the lipofectamine method (see above). Confocal images were obtained on a Zeiss CLSM 510, equipped with an argon laser (excitation, 488 nm) for the visualization of GFP and DsRed.

Results

Cloning of Kv6.3, Kv10.1, and Kv11.1. A search of the GenBank high throughput genomic sequence (*htgs*) database revealed genomic contigs containing exons coding for three previously uncharacterized homologues of Kv channels. The sequences of the genomic contigs were analyzed using GENEFINDER to determine the full coding sequences of the genes. The predicted proteins displayed the typical topology of a Kv subunit: six transmembrane segments (S1–S6), with an array of five to six positive charges in S4, and the potassium selectivity motif "GYG" in the P-loop between S5 and S6 (Fig. 1). Each gene was predicted to consist of two coding exons, without evidence for alternate splicing.

One of the predicted proteins consisted of 519 aa and shared more than 70% sequence identity with Kv6.1 and Kv6.2 (Fig. 1). Therefore, this protein has to be regarded as a previously uncharacterized member of the Kv6 subfamily, Kv6.3 or KCNG3. The other two proteins were composed of 436 and 545 aa and shared only ≈40% sequence identity with any of the previously identified Kv subunits. Therefore, we classified them as the first members of two previously uncharacterized subfamilies, Kv10.1 and Kv11.1.

The chromosomal locations of the genomic contigs containing Kv6.3, Kv10.1, and Kv11.1 are 16q24.1, 2p21, and 9p24.2, respectively. The complete cDNA sequences have been submitted to the GenBank database under accession nos. AF348982, AF348983, and AF348984 for Kv10.1, Kv11.1, and Kv6.3, respectively.

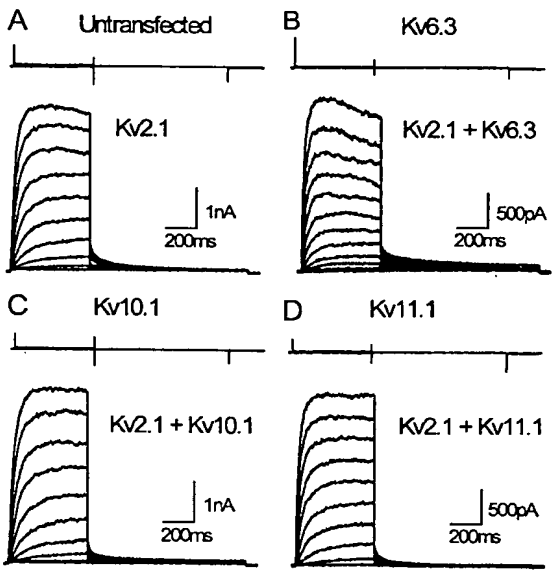


Fig. 3. Whole-cell current recordings of Kv6.3, Kv10.1, and Kv11.1, and the cotransfections with Kv2.1. The top sections in each panel show typical recordings for untransfected Ltk⁻ cells (*A*), or for cells expressing Kv6.3 (*B*), Kv10.1 (*C*), and Kv11.1 (*D*). The holding potential was −80 mV and cells were depolarized in 20-mV increments from −80 mV to +60 mV, 500 ms in duration, followed by a repolarizing pulse at −25 mV, 850 ms in duration. The bottom sections of each panel show typical recordings from Ltk⁻ cells expressing Kv2.1 (*A*), Kv2.1 + Kv6.3 (*B*), Kv2.1 + Kv10.1 (*C*), and Kv2.1 + Kv11.1 (*D*). The holding potential was −80 mV and cells were depolarized by 10-mV increments from −60 mV to +70 mV, 500 ms in duration. Deactivating tails were recorded at −25 mV or −35 mV for 850 ms.

Tissue Distribution of Kv6.3, Kv10.1, and Kv11.1. The search of the GenBank EST database yielded several hits for all three sequences, indicating that their mRNAs are indeed expressed. PCR analysis was used to assess the expression of Kv6.3, Kv10.1, and Kv11.1 mRNA in various human tissues (Fig. 2). Kv6.3 showed strong expression in brain and low expression in liver, small intestine, and colon. Kv10.1 was strongly expressed in pancreas and testis and weakly in brain, lung, kidney, thymus, ovary, small intestine, and colon. Kv11.1 gave a strong signal in pancreas and testis and a weaker signal in lung, liver, kidney, spleen, thymus, prostate, and ovary.

Functional Expression of Kv6.3, Kv10.1, and Kv11.1 in Ltk⁻ Cells. The coding sequences of Kv6.3, Kv10.1, and Kv11.1 were cloned into mammalian expression vectors for transient transfection in Ltk⁻ cells. The subunits Kv6.3, Kv10.1, and Kv11.1 each failed to generate current above background in these cells, as shown in the top sections of each panel in Fig. 3 ($n > 10$ cells, for at least two independent transfections for each clone). Previously discovered silent subunits can form heterotetrameric channels with the Kv2 subfamily (6–10). To test whether this could also be the case for the previously uncharacterized subunits, we performed coexpressions with Kv2.1.

Expression of the human Kv2.1 subunit alone resulted in a typical rapidly activating delayed outward rectifier K^+ current with functional properties as described (14, 15). The bottom sections of each panel in Fig. 3 show that coexpression with either previously uncharacterized subunit resulted in currents with distinct properties. For the cotransfection of Kv2.1 with Kv6.3, the threshold for activation was shifted by approximately 20 mV in hyperpolarizing direction compared with Kv2.1 alone

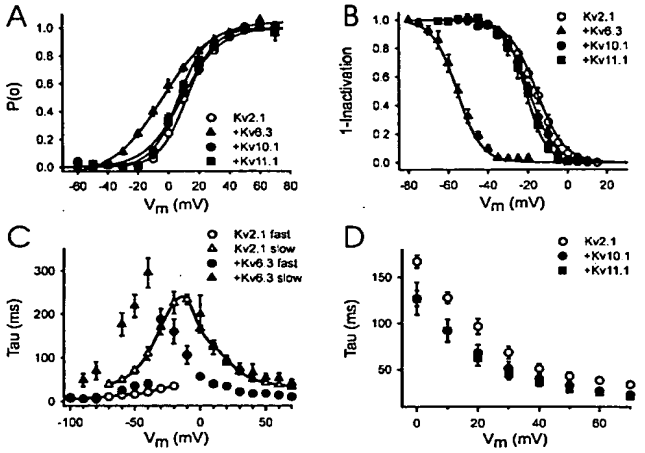


Fig. 4. (*A*) Voltage dependence of activation. The activation curves of Kv2.1, Kv2.1 + Kv6.3, Kv2.1 + Kv10.1, and Kv2.1 + Kv11.1 were obtained from the normalized initial tail amplitude recorded at −25 mV for Kv2.1, Kv2.1 + Kv10.1, and Kv2.1 + Kv11.1 or at −50 mV for Kv2.1 + Kv6.3 after 500-ms prepulses ranging from −60 mV to 70 mV in 10-mV steps. The solid line represents the Boltzmann function fitted to the experimental data (see *Experimental Procedures*). (*B*) Voltage dependence of inactivation. The inactivation curves of Kv2.1, Kv2.1 + Kv6.3, Kv2.1 + Kv10.1, and Kv2.1 + Kv11.1 were obtained from the normalized peak currents recorded during a 250-ms test pulse to 50 mV as a function of the 5-s prepulse ranging from −50 mV to 10 mV for Kv2.1, Kv2.1 + Kv10.1, and Kv2.1 + Kv11.1 and from −80 mV to −20 mV for Kv2.1 + Kv6.3. Experimental data were fitted with a Boltzmann function (solid lines). (*C*) Kinetics of activation and deactivation of Kv2.1 and Kv2.1 + Kv6.3. Mean time constants ± SEM of activation and deactivation are plotted as a function of the test potential. To obtain the time constants for activation, test pulses were applied ranging from −10 mV to 70 mV for Kv2.1 and −30 mV to 70 mV for Kv2.1 + Kv6.3 in 10-mV steps, 500 ms in duration. To obtain the time constants for deactivation, a 200-ms prepulse to 50 mV was followed by test pulses ranging from −20 mV to −100 mV in 10-mV steps, 850 ms in duration. The experimental data were fitted with mono- or double-exponential functions, as appropriate. The slow component of activation and deactivation are shown as triangles, whereas the fast components are shown as circles. WT Kv2.1 gating kinetics are connected with a solid line. (*D*) Kinetics of activation of Kv2.1, Kv2.1 + Kv10.1, and Kv2.1 + Kv11.1. Mean time constants ± SEM of activation are shown as a function of the step potentials (−10 mV to 70 mV). The pulse protocol for Kv2.1 + Kv10.1 and Kv2.1 + Kv11.1 is the same as for Kv2.1 alone in *C*.

(Fig. 4*A*). In addition, $V_{½}$ (Table 1) was significantly ($P < 0.001$) shifted toward hyperpolarizing voltages, and the slope decreased as well ($P < 0.001$). Cotransfection of Kv2.1 with Kv10.1 had no significant ($P > 0.05$) effect on the activation curve, whereas with Kv11.1 a small but consistent ($P < 0.05$) −5 mV shift was observed.

Co-transfection of Kv2.1 with Kv6.3 also markedly changed the C-type inactivation (Fig. 4*B*): the cotransfection resulted in a 40-mV hyperpolarizing shift in the voltage dependence of inactivation ($P < 0.001$). Cotransfection of Kv10.1 had no significant effect on the voltage dependence of inactivation ($P > 0.05$), whereas Kv11.1 gave a small −5 mV shift ($P < 0.05$).

The time-course of activation of Kv2.1 was fitted with a monoexponential function and resulted in time constants shown in Fig. 4 *C* and *D*. Upon cotransfection with Kv6.3, activation was accelerated (*P* values at all voltages <0.001) and the time course of activation was approximated better with a double exponential function (Fig. 4*C*). The acceleration of activation was less pronounced for the cotransfection of Kv10.1 or Kv11.1 (Fig. 4*D*), but still statistically significant ($P < 0.05$ at all voltages). Deactivation was fitted with a mono- or double-exponential function as appropriate. Cotransfection of Kv6.3 slowed deactivation significantly (*P* value at all voltages <0.001), whereas Kv10.1 and Kv11.1 had no significant effect (*P* value at all voltages >0.05). A summary of the electrophysiological parameters is given in Table 1.

T \ B	Kv1.5	Kv2.1	Kv3.1	Kv4.3	Kv5.1	Kv6.1	Kv6.3	Kv8.1	Kv9.3	Kv10.1	Kv11.1
Kv6.3	-	+	+	-	+	-	-	-	-	-	-
Kv10.1	-	+	+	-	+	-	-	-	-	-	-
Kv11.1	-	+	+	-	+	-	-	-	-	-	-
Kv2.1	-	+									

Fig. 5. Interaction of Kv6.3, Kv10.1, and Kv11.1 with representative subunits of all Kv subfamilies. The intracellular N-terminal segment that contains the subfamily-specific NAB domain was used as bait (B) and/or target (T) in a yeast two-hybrid analysis.

Tissue Distribution of Kv2.1. To test whether the previously uncharacterized subunits could be regulatory subunits for Kv2.1 *in vivo*, we determined the expression of Kv2.1 with the same cDNA panels as we did for Kv6.3, Kv10.1, and Kv11.1 (Fig. 2). Kv2.1 showed very high expression in brain, skeletal muscle, pancreas, and small intestine and moderate to high expression in heart, placenta, lung, liver, kidney, spleen, thymus, prostate, testis, ovary, and colon, consistent with previous reports (6, 16, 17). These results show that Kv6.3, Kv10.1, and Kv11.1 are expressed in several tissues in which Kv2.1 is also expressed, indicating that at least in some tissues these subunits could indeed interact with Kv2.1 to form heterotetrameric channels.

Biochemical Evidence for Selective Interaction with Kv Subunits. To explore in a more unbiased manner the potential interactions of the three previously uncharacterized subunits with subunits of the known subfamilies, we used a yeast two-hybrid approach. Given the limitations of this method, we screened with the intracellular amino terminal segment, which contains the NAB domain that regulates coassembly (18–21). Kv6.3, Kv10.1, and Kv11.1 each did not show interactions with themselves, nor with each other, Kv1.5, Kv4.3, Kv8.1, and Kv9.1 (Fig. 5). In contrast, a strong interaction with Kv2.1, Kv3.1, and Kv5.1 was seen. For each of the previously uncharacterized subunits this interaction was as strong as the interaction of Kv2.1 with itself (positive

Table 1. Electrophysiological parameters

	Voltage dependence						Time constants, ms					
	Activation			Inactivation			Activation at 0 mV			Deactivation at −40 mV		
	$V_{1/2}$, mV	*k*, mV	*n*	$V_{1/2}$, mV	*k*, mV	*n*	Fast	Slow	*n*	Fast	Slow	*n*
Kv2.1	12.2 ± 1.4	9.5 ± 0.6	11	−15.9 ± 1.2	7.2 ± 0.6	5	167 ± 12	N.A.	9	21.4 ± 2.1	110 ± 15	5
+ Kv6.3	−4.2 ± 0.7	15.1 ± 0.7	5	−55.6 ± 1.1	6.6 ± 0.8	6	58.2 ± 4.9	200 ± 38	8	40.1 ± 9.2	295 ± 34	5
+ Kv10.1	9.3 ± 2.0	9.8 ± 0.7	9	−19.8 ± 1.2	6.4 ± 0.2	5	127 ± 8	N.A.	9	18.0 ± 1.2	115 ± 6	6
+ Kv11.1	7.0 ± 1.3	9.8 ± 0.7	11	−21.2 ± 1.7	5.2 ± 0.2	7	127 ± 18	N.A.	10	18.8 ± 1.2	89.6 ± 3.1	7

Values are given as mean ± SEM; *n* = number of experiments. $V_{1/2}$ and k obtained from Boltzmann fit (see *Experimental Procedures*). N.A., not applicable.

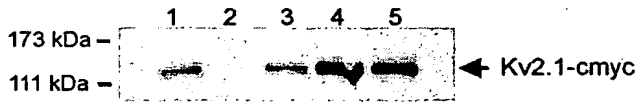


Fig. 6. Co-immunoprecipitation of Kv6.3GFP, Kv10.1GFP, and Kv11.1GFP with Kv2.1c-myc. Immunoprecipitation was done with anti-GFP antibodies, Western blot was performed with anti-c-myc. Lanes 3–5 show that Kv2.1c-myc was coprecipitated with Kv6.3GFP, Kv10.1GFP, and Kv11.1GFP. GFP-tagged Kv2.1 (lane 1) and Kv1.5 (lane2) were used as positive and negative controls, respectively.

control). Kv2.1 failed to interact with Kv1.5, consistent with the known lack of heterotetramerization between Kv1.5 and Kv2.1 (18, 22). The interaction of Kv2.1 with the previously uncharacterized subunits was confirmed with coimmunoprecipitation, using the full-length proteins (Fig. 6).

Subcellular Localization of Kv6.3, Kv10.1, and Kv11.1. Although the lack of N-terminal tetramerization might explain the lack of current, it is known that Kv2.1 can generate current when the NAB domain is removed (18). Therefore, we also determined the subcellular localization of the previously uncharacterized subunits by using confocal microscopy. To visualize the subcellular protein distribution, GFP was fused to their carboxy termini. Transfected cells expressing only Kv6.3, Kv10.1, or Kv11.1 showed a punctated intracellular appearance without staining of the plasma membrane (Fig. 7, column 1). This indicates that the full-length protein was made, because GFP was added on the C-terminal end. To test whether this pattern reflected retention in the ER, we performed cotransfections with a vector (DsRed-ER) containing the cDNA from the red fluorescent protein DsRed, fused with the ER targeting signal from calreticulin and the ER retention signal, KDEL (Fig. 7, column 2). The localization of the red and green fluorescence overlapped completely, resulting in a yellow-orange color indicating that each of the three subunits were retained in the ER when they were expressed alone (Fig. 7, column 3). When Kv2.1 was coexpressed with these subunits, a redistribution of the green fluorescence was observed. In each case, prominent GFP staining was evident at the plasma membrane with minimal intracellular staining (Fig. 7, column 4). Potassium currents obtained with the GFP-tagged subunits were similar to those shown in Figs. 3 and 4. The intracellular staining was a nearly pure DsRed-ER fluorescence, showing hardly any overlap (Fig. 7, column 5). These results indicate that Kv2.1 promotes trafficking of Kv6.3, Kv10.1, and Kv11.1 to the cell surface membrane, presumably by forming heterotetrameric channels with these subunits.

Discussion

This study reports the cloning and characterization of three previously uncharacterized α-subunits of voltage-gated potassium channels: Kv6.3, Kv10.1, and Kv11.1. The conventional methods to clone potassium channels include homology and expression cloning (14, 23). The disadvantage of both techniques is their dependence on expression level or on a functional signature: genes with very low expression levels or lacking a functional signature are not (easily) picked up by these techniques. The human genome project does allow to detect and clone such genes, as is demonstrated here for Kv6.3, Kv10.1, and Kv11.1.

When expressed in mammalian Ltk$^-$ cells each of the three subunits was unable to elicit any current, indicating that they belong to the "silent" subunits (6–10). The lack of functional current can be explained by retention in the ER, as was demonstrated with confocal microscopy, comparable with observations for Kv8 and Kv9 subunits (9, 16, 24). Such retention can have various causes such as ER retention signals or improper folding and/or assembly. Investigation of the sequences of Kv6.3, Kv10.1, and Kv11.1 did not reveal known ER retention or export signals, suggesting an assembly problem. For the confocal imaging we used a C-terminal GFP tag, which could interfere with trafficking. Indeed, C-terminal sequences can control efficient cell surface expression and clustering (25, 26). However, the three subunits reported here do not display such sequences and the GFP tag did not effect the currents recorded after coexpression. Inefficient assembly of channel subunits might originate from the aminoterminal "NAB" domain that directs and restricts subunit assembly within Kv subfamilies (18–21). Indeed, the aminotermini of Kv6.3, Kv10.1, and Kv11.1 did not interact with themselves, as was demonstrated with a yeast two-hybrid analysis. Therefore, to the extent that the NAB domain facilitates homotetrameric assembly, these subunits would appear incapable of efficient homotetramerization, which might explain ER retention.

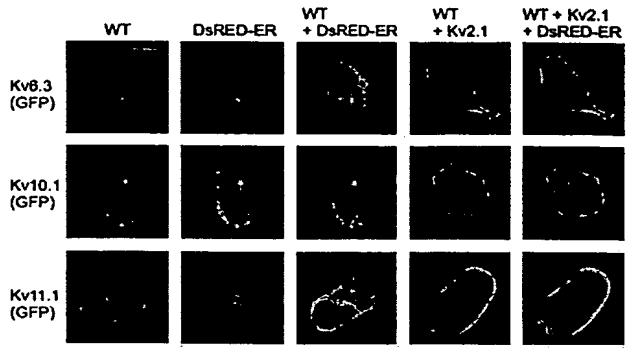


Fig. 7. Subcellular localization of the channel-GFP fusion proteins assessed by confocal imaging. The rows show images with GFP fusion proteins of Kv6.3, Kv10.1, and Kv11.1, respectively. The first three columns show the fluorescence of the channel subunits, the DsRed-ER localization vector, and the overlay of both, respectively. The last two columns show cells cotransfected with Kv2.1, DsRed-ER, and each of the subunits; the surface staining of the GFP-tagged subunits (fourth column) is obvious with minimal overlap with the DsRed-ER fluorescence (overlay of both in the fifth column). (Scale bar, 10 μm.)

However, these incompatible amino termini may not be the only reason for the lack of functionality for these and other silent Kv channels. Indeed, distinct currents were observed for a chimera between the N terminus of Kv8.1 in a Kv1.3 background (7). However, a chimera with S6 from Kv8.1 in a Kv1.3 background (and *vice versa*) was not functional, which indicates that part of the nonfunctionality resides in the S6 segment. The alignment of this segment (Fig. 8) demonstrates that the three subunits reported here, as well as previously cloned silent subunits, all lack the second proline of the conserved P–X–P motif of the Kv1–Kv4 subunits. This points to a major structural difference in the S6 segments between the functional and silent

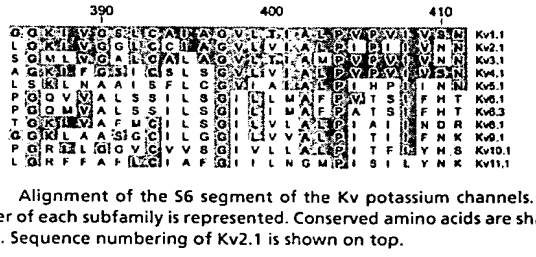


Fig. 8. Alignment of the S6 segment of the Kv potassium channels. One member of each subfamily is represented. Conserved amino acids are shaded in gray. Sequence numbering of Kv2.1 is shown on top.

subunits. However, when P406 and V409 of Kv2.1 were mutated to the corresponding residues of Kv8.1, altered but functional currents were observed, indicating that these residues alone do not explain the nonfunctional S6 chimera (24). However, P406 is the second proline in the highly conserved P–X–P motif from the functional Kv channels and might be responsible for a sharp bend in the S6 helical structure involved in gating (27). All of the silent subunits lack the second proline of the P–X–P motif, indicating a structural difference of the S6 segment between the functional and the silent subunits, which is apparently compensated in the heterotetrameric configuration.

The profound effects of Kv6.3 on Kv2.1 gating properties suggest an important role for these heterotetramers: the latter would be inactivated at potentials close to resting potential ($V_{½}$ for inactivation is −56 mV) in contrast to the homotetrameric Kv2.1 channels ($V_{½}$ = −16 mV). Because both subunits are expressed in the brain (Fig. 2) functional heterotetramers could exist (6, 17). Previous studies on the sustained delayed rectifier component of hippocampal neurons showed properties that are comparable with those of Kv2.1 and Kv6.3 heteromultimers (28, 29). At −5 mV the two time constants for activation for the current in those neurons were 53 ms and 190 ms, which is comparable with heterotetrameric channels of Kv2.1 and Kv6.3 (Table 1). In addition, the midpoint of inactivation was more negative (−96 mV), which is at least closer to −56 mV for Kv2.1 and Kv6.3 compared with −16 mV for Kv2.1 alone. Furthermore, the pharmacological profile for homomeric Kv2.1 channels did not correspond completely with that of the sustained delayed rectifier component: the TEA sensitivity depended on the cell type under investigation (29) and differed from Kv2.1. Thus native channels that are considered to contain Kv2 subunits may well be heterotetramers, although it will be a challenge to assign the proper heterotetrameric combination.

In neurons, Kv2.1 is thought to have a role in controlling the membrane potential and in the electrical signaling of cells (30, 31). Using antisense oligonucleotides it was demonstrated that somato-dendritic excitability was regulated by Kv2.1 in hippocampal neurons (32). The down-regulation of the Kv2.1 protein (>90%) was associated with action potential broadening and an increase in intracellular calcium at high-frequency stimulation. The gating properties were not reported in this study but the 90% down regulation of the Kv2.1 protein was associated with only a 50% reduction of the sustained delayed rectifier component. Although the molecular nature of the sustained current remains to be elucidated, a heterotetrameric subunit composition could be compatible with such dominant negative results.

Within the Kv1, Kv3, and Kv4 families, functional diversity is achieved by the different properties of each subunit, and by the heteromeric (intrafamily) assembly of α-subunits resulting in channels with distinct biophysical properties. The Kv2 subfamily contains only two members that have very similar biophysical properties. While Kv2.1 and Kv2.2 are capable of heteromultimerization, the resulting currents are functionally similar to those of their homotetramers (33). It has been suggested that the functional diversity within this family is achieved through heteromeric assembly with other subfamilies of silent subunits (34). Our results now add three more subunits that can expand further the functional diversity in different types of tissue or during development.

Thus far, 19 functional Kv α-subunits had been discovered and only 7 silent subunits. Our results enlarge this last group to 10 subunits. Despite the large number of these subunits, their exact physiological role is still poorly understood mainly because of the difficulty in recognizing a silent subunit in isolated cells or in tissue. Thus far, heterologous expression studies have led to the hypothesis that the silent subunits must interact with other Kv subunits from the Kv2 and Kv3 subfamilies to regulate their function. If each of the silent subunits can interact with the two members of the Kv2 subfamily and the four members of the Kv3 subfamily, then at least 60 different heterotetramers are possible (each with one to three silent subunits). Thus, this growing group of silent subunits considerably expands the potential for molecular diversity of the native K^+ channels. Thus, future experiments will be necessary to reveal the true interaction partners and the physiological importance of the silent subunits.

Note Added in Proof. While this paper was under review, another group (35) reported cloning of "Kv6.3," which corresponds to Kv10.1 in our analysis (see Fig. 1).

We thank Dr. Jean-Pierre Timmermans for the use of the confocal microscope. This work was supported by Flanders Institute for Biotechnology Grant PRJ05 and National Institutes of Health/National Heart, Lung, and Blood Institute Grant HL59689.

1. Hille, B. (1991) *Ionic Channels of Excitable Membranes* (Sinauer, Sunderland, MA).
2. Barry, D. M. & Nerbonne, J. M. (1996) *Annu. Rev. Physiol.* **58**, 363–394.
3. Pongs, O. (1999) *FEBS Lett.* **452**, 31–35.
4. Pongs, O. (1992) *Physiol. Rev.* **72**, S69–S88.
5. Zhao, B., Rassendren, F., Kaang, B. K., Furukawa, Y., Kubo, T. & Kandel, E. R. (1994) *Neuron* **13**, 1205–1213.
6. Drewe, J. A., Verma, S., Frech, G. & Joho, R. H. (1992) *J. Neurosci.* **12**, 538–548.
7. Hugnot, J. P., Salinas, M., Lesage, F., Guillemare, E., de Weille, J., Heurteaux, C., Mattei, M. G. & Lazdunski, M. (1996) *EMBO J.* **15**, 3322–3331.
8. Patel, A. J., Lazdunski, M. & Honore, E. (1997) *EMBO J.* **16**, 6615–6625.
9. Salinas, M., Duprat, F., Heurteaux, C., Hugnot, J. P. & Lazdunski, M. (1997) *J. Biol. Chem.* **272**, 24371–24379.
10. Zhu, X. R., Netzer, R., Bohlke, K., Liu, Q. & Pongs, O. (1999) *Receptors Channels* **6**, 337–350.
11. Post, M. A., Kirsch, G. E. & Brown, A. M. (1996) *FEBS Lett.* **399**, 177–182.
12. Snyders, D. J. & Chaudhary, A. C. (1996) *Mol. Pharmacol.* **49**, 949–955.
13. Hamill, O. P., Marty, A., Neher, E., Sakmann, B. & Sigworth, F. J. (1981) *Pflügers Arch. Eur. J. Physiol.* **391**, 85–100.
14. Frech, G. C., VanDongen, A. M., Schuster, G., Brown, A. M. & Joho, R. H. (1989) *Nature (London)* **340**, 642–645.
15. Benndorf, K., Koopmann, R., Lorra, C. & Pongs, O. (1994) *J. Physiol. (London)* **477**, 1–14.
16. Shepard, A. R. & Rae, J. L. (1999) *Am. J. Physiol.* **277**, C412–C424.
17. Trimmer, J. S. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 10764–10768.
18. Li, M., Jan, Y. N. & Jan, L. Y. (1992) *Science* **257**, 1225–1230.
19. Xu, J., Yu, W., Jan, Y. N., Jan, L. Y. & Li, M. (1995) *J. Biol. Chem.* **270**, 24761–24768.
20. Papazian, D. M. (1999) *Neuron* **23**, 7–10.
21. Shen, N. V. & Pfaffinger, P. J. (1995) *Neuron* **14**, 625–633.
22. Covarrubias, M., Wei, A. A. & Salkoff, L. (1991) *Neuron* **7**, 763–773.
23. Tamkun, M. M., Knoth, K. M., Walbridge, J. A., Kroemer, H., Roden, D. M. & Glover, D. M. (1991) *FASEB J.* **5**, 331–337.
24. Salinas, M., de Weille, J., Guillemare, E., Lazdunski, M. & Hugnot, J. P. (1997) *J. Biol. Chem.* **272**, 8774–8780.
25. Burke, N. A., Takimoto, K., Li, D., Han, W., Watkins, S. C. & Levitan, E. S. (1999) *J. Gen. Physiol.* **113**, 71–80.
26. Li, D., Takimoto, K. & Levitan, E. S. (2000) *J. Biol. Chem.* **275**, 11597–11602.
27. del Camino, D., Holmgren, M., Liu, Y. & Yellen, G. (2000) *Nature (London)* **403**, 321–325.
28. Numann, R. E., Wadman, W. J. & Wong, R. K. (1987) *J. Physiol. (London)* **393**, 331–353.
29. Zhang, L. & McBain, C. J. (1995) *J. Physiol. (London)* **488**, 647–660.
30. Murakoshi, H. & Trimmer, J. S. (1999) *J. Neurosci.* **19**, 1728–1735.
31. Baranauskas, G., Tkatch, T. & Surmeier, D. J. (1999) *J. Neurosci.* **19**, 6394–6404.
32. Du, J., Haak, L. L., Phillips, T. E., Russell, J. T. & McBain, C. J. (2000) *J. Physiol. (London)* **522**, 19–31.
33. Blaine, J. T. & Ribera, A. B. (1998) *J. Neurosci.* **18**, 9585–9593.
34. Kramer, J. W., Post, M. A., Brown, A. M. & Kirsch, G. E. (1998) *Am. J. Physiol.* **274**, C1501–C1510.
35. Sano, Y., Mochizuki, S., Miyake, A., Kitada, C., Inamura, K., Yokoi, H., Nozawa, K., Matsushime, H. & Furuichi, K. (2002) *FEBS Lett.* **512**, 230–234.

EXHIBIT F

1: Brain Res Mol Brain Res. 2004 Apr 7;123(1-2):91-103.

Modification of Kv2.1 K+ currents by the silent Kv10 subunits.

Vega-Saenz de Miera EC.

Department of Physiology and Neuroscience, New York University School of Medicine, 550 First Avenue, New York, NY 10016, USA.
vegae01@endeavor.med.nyu.edu

Human and rat Kv10.1a and b cDNAs encode silent K+ channel pore-forming subunits that modify the electrophysiological properties of Kv2.1. These alternatively spliced variants arise by the usage of an alternative site of splicing in exon 1 producing an 11-amino acid insertion in the linker between the first and second transmembrane domains in Kv10.1b. In human, the Kv10s mRNA were detected by Northern blot in brain kidney lung and pancreas. In brain, they were expressed in cortex, hippocampus, caudate, putamen, amygdala and weakly in substantia nigra. In rat, Kv10.1 products were detected in brain and weakly in testes. In situ hybridization in rat brain shows that Kv10.1 mRNAs are expressed in cortex, olfactory cortical structures, basal ganglia/striatal structures, hippocampus and in many nuclei of the amygdala complex. The CA3 and dentate gyrus of the hippocampus present a gradient that show a progression from high level of expression in the caudo-ventro-medial area to a weak level in the dorso-rostral area. The CA1 and CA2 areas had low levels throughout the hippocampus. Several small nuclei were also labeled in the thalamus, hypothalamus, pons, midbrain, and medulla oblongata. Co-injection of Kv2.1 and Kv10.1a or b mRNAs in Xenopus oocytes produced smaller currents that in the Kv2.1 injected oocytes and a moderate reduction of the inactivation rate without any appreciable change in recovery from inactivation or voltage dependence of activation or inactivation. At higher concentration, Kv10.1a also reduces the activation rate and a more important reduction in the inactivation rate. The gene that encodes for Kv10.1 mRNAs maps to chromosome 2p22.1 in human, 6q12 in rat and 17E4 in mouse, locations consistent with the known systeny for human, rat and mouse chromosomes.

MeSH Terms:

- Alternative Splicing/genetics
- Amino Acid Sequence/genetics

- Animals
- Base Sequence/genetics
- Brain/metabolism*
- Brain Chemistry/genetics*
- Chromosomes, Human, Pair 2/genetics
- DNA, Complementary/analysis
- DNA, Complementary/genetics
- Humans
- Membrane Potentials/genetics
- Mice
- Molecular Sequence Data
- Oocytes/metabolism
- Phylogeny
- Potassium Channels/genetics*
- Potassium Channels/metabolism
- Potassium Channels, Voltage-Gated*
- Protein Isoforms/genetics
- Protein Isoforms/metabolism
- Protein Subunits/genetics
- Protein Subunits/metabolism
- RNA, Messenger/metabolism
- Rats
- Research Support, Non-U.S. Gov't
- Sequence Homology, Amino Acid
- Sequence Homology, Nucleic Acid
- Viscera/metabolism
- Xenopus

Substances:

- DNA, Complementary
- Kv6.3 protein, mouse
- Potassium Channels
- Potassium Channels, Voltage-Gated
- Protein Isoforms
- Protein Subunits
- RNA, Messenger
- delayed rectifier potassium channel

Secondary Source ID:

- GENBANK/AF454547
- GENBANK/AF454548
- GENBANK/AF454549
- GENBANK/AF454550
- GENBANK/AF454551
- GENBANK/AF454552

PMID: 15046870 [PubMed - indexed for MEDLINE]

Feb 25 2005 07:07:33

EXHIBIT G

1: Adv Exp Med Biol. 2001;502:401-18.

Gene transfer and metabolic modulators as new therapies for pulmonary hypertension. Increasing expression and activity of potassium channels in rat and human models.

Michelakis ED, Dyck JR, McMurtry MS, Wang S, Wu XC, Moudgil R, Hashimoto K, Puttagunta L, Archer SL.

Department of Medicine (Cardiology), University of Alberta, Edmonton, Canada.

Chronic Hypoxic Pulmonary Hypertension (CH-PHT) is characterized by pulmonary artery (PA) vasoconstriction and cell proliferation/hypertrophy. PA smooth muscle cell (PASMC) contractility and proliferation are controlled by cytosolic Ca++ levels, which are largely determined by membrane potential (E (M)). E(M) is depolarized in CH-PHT due to decreased expression and functional inhibition of several redox-regulated, 4-aminopyridine (4-AP) sensitive, voltage-gated K+ channels (Kv1.5 and Kv2.1). Humans with Pulmonary Arterial Hypertension (PAH) also have decreased PASMC expression of Kv1.5 and Kv2.1. We speculate this "K+-channelopathy" contributes to PASMC depolarization and Ca++ overload thus promoting vasoconstriction and PASMC proliferation. We hypothesized that restoration of Kv channel expression in PHT and might eventually be beneficial. METHODS: Two strategies were used to increase Kv channel expression in PASMCs: oral administration of a metabolic modulator drug (Dichloroacetate, DCA) and direct Kv gene transfer using an adenovirus (Ad5-Kv2.1). DCA a pyruvate dehydrogenase kinase inhibitor, promotes a more oxidized redox state mimicking normoxia and previously has been noted to increase K+ current in myocytes. Rats were given DCA in the drinking water after the development of CH-PHT and hemodynamics were measured approximately 5 days later. We also tested the ability of Ad5-Kv2.1 to increase Kv2.1 channel expression and function in human PAs ex vivo. RESULTS: The DCA-treated rats had decreased PVR, RVH and PA remodeling compared to the control CH-PHT rats (n=5/group, $p<0.05$). DCA restored Kv2.1 expression and PASMC Kv current density to near normoxic levels. Adenoviral gene transfer increased expression of Kv2.1 channels and enhanced 4-AP constriction in human PAs. CONCLUSION: Increasing Kv channel function in PAs is feasible and might

be beneficial.

PMID: 11950153 [PubMed - indexed for MEDLINE]

Downregulation of voltage-gated K^+ channels in rat heart with right ventricular hypertrophy

JONG-KOOK LEE,[1] ATSUSHI NISHIYAMA,[1] FUKUSHI KAMBE,[2] HISAO SEO,[2] SUSUMU TAKEUCHI,[1] KAICHIRO KAMIYA,[1] ITSUO KODAMA,[1] AND JUNJI TOYAMA[1]
Departments of [1]Circulation and [2]Endocrinology and Metabolism, Research Institute of Environmental Medicine, Nagoya University, Nagoya, Japan 464-8601

Lee, Jong-Kook, Atsushi Nishiyama, Fukushi Kambe, Hisao Seo, Susumu Takeuchi, Kaichiro Kamiya, Itsuo Kodama, and Junji Toyama. Downregulation of voltage-gated K^+ channels in rat heart with right ventricular hypertrophy. *Am. J. Physiol.* 277 (*Heart Circ. Physiol.* 46): H1725–H1731, 1999.—The effects of myocardial hypertrophy on mRNA expression levels of voltage-gated K^+ channels were investigated using monocrotaline (MCT)-induced pulmonary hypertensive rats. The ratio of right ventricle weight to left ventricle plus septum weight on *day 28* was increased significantly compared with control rats [control vs. MCT: 0.27 ± 0.01 vs. 0.58 ± 0.03 ms (n = 8–13); $P < 0.05$]. Electrocardiograms showed that QRS duration [control vs. MCT: 26.4 ± 2.6 ms vs. 31.5 ± 5.8 ms (n = 6); $P < 0.05$], Q-T interval [control vs. MCT: 100.8 ± 8.9 ms vs. 110.0 ± 4.2 ms (n = 6); $P < 0.05$] and corrected Q-T interval [Q-T_c; control vs. MCT: 8.4 ± 0.7 ms vs. 10.2 ± 0.4 ms (n = 6); $P < 0.05$] were prolonged significantly on *day 28*. mRNA levels of Kv1.2, 1.5, 2.1, 4.2, and 4.3 for *day 28* assessed by ribonuclease protection assays were decreased significantly from control by 60 ± 10, 76 ± 3, 58 ± 5, 81 ± 5, and 45 ± 12%, respectively (n = 3; $P < 0.005$), and Kv1.4 mRNA level for *day 28* was unaffected [Kv1.4, control vs. MCT: 1.0 ± 0.28 vs. 0.88 ± 0.44 (arbitrary units) (n = 3); not significant (NS)]. On the other hand, there was no significant difference between control and MCT rats in mRNA levels of these Kv channels for *day 14* [Kv1.2 (control vs. MCT): 1.0 ± 0.25 vs. 0.87 ± 0.18 (n = 3), NS; Kv1.4: 1.0 ± 0.22 vs. 1.27 ± 0.37 (n = 3), NS; Kv1.5: 1.0 ± 0.16 vs. 0.91 ± 0.28 (n = 3), NS; Kv2.1: 1.0 ± 0.26 vs. 0.99 ± 0.25 (n = 3), NS; Kv4.2: 1.0 ± 0.15 vs. 1.22 ± 0.28 (n = 3), NS; Kv4.3: 1.0 ± 0.20 vs. 1.21 ± 0.28 (n = 3), NS]. These findings suggest that altered ventricular repolarization at the advanced stage of hypertrophy may be the result of an inhibition of gene expression of multiple types of voltage-gated K^+ channels.

ventricular hypertrophy; voltage-gated potassium channels; messenger ribonucleic acid expression

CLINICAL STUDIES have suggested that ventricular hypertrophy is associated with a greater risk of sudden cardiac death probably caused by lethal ventricular arrhythmias (18). Alterations of repolarization are often recognized in clinical electrocardiograms (ECGs) with the development of ventricular hypertrophy. Cellular electrophysiological studies have shown that these alterations in repolarization are caused by the prolongation of action potential duration (APD) (1). APD prolongation has been ascribed to a decrease of the transient outward K^+ current (I_{to}) density or an increase of the L-type Ca^{2+} current (I_{Ca}) density in a variety of experimental models of cardiac hypertrophy in rats (5, 7, 15, 27, 31), cats (12, 16, 24), and guinea pigs (26). Comparable changes in I_{to} and I_{Ca} were also reported in human patients with an advanced stage of congestive heart failure (6).

In a recent study using rats with monocrotaline (MCT)-induced right ventricular (RV) hypertrophy, we reported (17) that hypertrophy was associated with stage-dependent changes in I_{to} and I_{Ca}; the APD prolongation in the early compensated stage of hypertrophy may be caused mainly by an increase of I_{Ca} density, whereas the APD prolongation in the advanced stage of hypertrophy may be the result of a reduction of I_{to} density. The decrease of I_{to} density at the late stage of hypertrophy is consistent with previous reports on other models of ventricular hypertrophy (5, 7, 31). In adult rat hearts, many voltage-gated K^+ channel subunits have been cloned, which include Kv1.2, Kv1.4, Kv1.5, Kv2.1, Kv4.2, and Kv4.3. Kv4.2 and Kv4.3 of the *Shal* family are the most likely candidates for I_{to} (3, 9), whereas Kv1.2 and Kv1.5 of the *Shaker* family and Kv2.1 of the *Shab* family are considered as candidates for other delayed rectifier K^+ channels sensitive to 4-aminopyridine (4-AP) or tetraethylammonium (TEA) (4, 8). It has been shown in several studies that the expression of these cloned K^+ channels is affected in certain pathophysiological conditions including cardiac hypertrophy and hormonal abnormalities (20, 23, 29, 30). The molecular mechanisms for altered repolarization in cardiac hypertrophy are, however, still unsettled and controversial (20, 30).

In the present study, we investigated changes in voltage-gated K^+ channel gene expression in hypertrophied rat hearts with MCT-induced pulmonary hypertension. mRNA levels of Kv1.2, Kv1.4, Kv1.5, Kv2.1, Kv4.2, and Kv4.3 α-subunits were measured by ribonuclease protection assay (RPA). The results have revealed that not only Kv4.2 and Kv4.3 but also Kv1.2, Kv1.5, and Kv2.1 mRNA are downregulated in the hypertrophied ventricle. Such alteration in multiple types of voltage-gated K^+ channel gene expression may contribute to the repolarization delay in an advanced stage of ventricular hypertrophy.

MATERIALS AND METHODS

Animals. Five-week-old male Wistar rats weighing 170–190 g were treated with MCT (Sigma, St. Louis, MO) to produce pulmonary hypertension as described previously (14, 17, 22). In brief, a single dose of 60 mg/kg MCT, which was dissolved in 1 N HCl neutralized with 0.5 N NaOH and diluted with sterile distilled water to obtain a 2% solution,

The costs of publication of this article were defrayed in part by the payment of page charges. The article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

was injected subcutaneously into the interscapular region. In control rats of corresponding age and weight, saline was injected instead of MCT. The rats were allowed to eat freely from a supply of standard rat chow. The animals were killed under ether anesthesia on the day of MCT or saline injection (*day 0*) or 7, 14, 21, and 28 days after the injection. The hearts were removed quickly and used for estimation of RV hypertrophy as well as for cell isolation and mRNA measurements. RV hypertrophy was estimated by measuring the ratio of the RV free wall tissue weight to body weight (BW) and that of RV weight to left ventricular free wall plus septum (LV+S) weight.

Electrocardiograms. ECGs were recorded immediately before and on *days 14* and *28* after the injection of MCT. Under anesthesia (20 mg/kg pentobarbital sodium ip), leads I and II were recorded. The signals were stored with a digital audio recording system (Sony, Tokyo, Japan), and the ECG parameters were analyzed using software (Softron, Tokyo, Japan) programmed for the analysis of ECG parameters in rodents.

Ribonuclease protection assay. For the RPA, rat Kv1.2, Kv1.4, Kv1.5, Kv2.1, Kv4.2, and Kv4.3 α-subunit cDNA fragments were amplified by RT-PCR. The nucleotide sequences of the primers and the amplified regions are described here. Nucleotide numbers for each primer correspond to those from the translation start site: Kv1.2: sense 5′-AAGCTTTAACTGATGTCTGATTGAAACCTA-3′, antisense 5′-GATGCTGGCTCCATGGGTGAC-3′, nucleotides 1,487–1,743 (21); Kv1.4: sense 5′-AAGCTTTCTACTTCTTCTTCCCTGGGGGAC-3′, antisense 5′-TGCATCACTTATTTGATATGC-3′, nucleotides 1,801–2,132 (32); Kv1.5: sense 5′-CCGAGTATTTAAGCCCACCTG-3′, antisense 5′-CTAAGCTTTTTAAAGTCAAATTTG-3′, nucleotides 1,888–2,144 (28); Kv2.1: sense 5′-AAGCTTGCTCTGGTTTCTTCGTGGAGAGTC-3′, antisense 5′-CACGCTGTAGAGCAGCTGACC-3′, nucleotides 1,931–2,295 (11); Kv4.2: sense 5′-TACCGCACGGGGAAGCTTCACTAT-3′, antisense 5′-TGGAACTGTTTCCACCACATTCGC-3′, nucleotides 295–624 (2); Kv4.3: sense 5′-AAGCTTGGCACCCCAGAAGAGGAGCATG-3′, antisense 5′-GTTGGAGTTGGGCAGGTGCGTGGT-3′, nucleotides 1,372–1,626 (10). A *Hind* III site (AAGCTT) was introduced into the 5′ end of the sense primers of the Kv1.2, Kv1.4, Kv2.1, and Kv4.3 (underlined). In Kv4.2, a *Hind* III site is present in the coding region (underlined). The amplified cDNA was cloned into pGEM-T vector using the TA cloning system (Promega, Madison, WI).

The plasmids containing cDNAs were linearized by digestion with an appropriate restriction enzyme (*Hind* III for Kv1.2, Kv1.4, Kv2.1, Kv4.2, and Kv4.3; *Nco*I present in pGEM-T vector for Kv1.5). Antisense cRNA probes were prepared using a MAXIscript kit (Ambion, Austin, TX) and [α-^{32}P]UTP (Du Pont-New England Nuclear). The cyclophilin cRNA probe was also prepared from the cDNA purchased from Ambion (pTRI-cyclophilin-rat antisense control template, nucleotides 38–142) to detect cyclophilin mRNA as an internal control. RPA was performed using a HybSpeed RPA kit (Ambion) according to the manufacturer's protocol. Hybridization of the two probes [2 × 10^4 counts/min (cpm) Kv4.2 cRNA and 2 × 10^4 cpm cyclophilin cRNA] with 10 μg total RNA was carried out at 68°C for 10 min, followed by digestion with RNase A and RNase T1 at 37°C for 30 min. The reaction was terminated by addition of sodium dodecyl sulfate and proteinase K, followed by phenol-chloroform extraction and ethanol precipitation. The protected fragments were visualized by autoradiography after electrophoresis on a 5% polyacrylamide/8 M urea gel. Quantitative analysis was carried out using Fujix Bioimage Analyzer with which we measured the radioactivity of the bands in a selected area. Each mRNA level of Kv channels was normalized by the levels of cyclophilin. The mRNA of each lane in the gels is from different animals.

Statistics. Data are expressed as means ± SE. Statistical analyses were performed using one-way analysis of variance with multiple comparisons. Differences were considered significant at $P < 0.05$.

RESULTS

Characteristics of experimental animals. Table 1 summarizes BW and heart, lung, liver, and kidney weight before and after the injection of MCT. There was no significant difference in BW between groups on *days 0* and *14*, but BW of MCT-treated rats were significantly decreased by 19.3% compared with those of control rats on *day 28*. The ratio of RV weight to BW and the ratio of RV weight to LV+S weight were both increased significantly on *days 14* and *28*, whereas the ratio of LV+S weight to BW was unaffected during the entire observation period. There was no significant difference in the ratio of kidney and liver weights to BW. On the other hand, there was a significant increase in the ratio of lung weight to BW in the MCT rats, probably because of the primary pathological effects of MCT on the lung. Eleven of twelve MCT rats showed the signs of right-sided heart failure during the following week, including tachypnea, ascites, pleural effusion, edematous extremities, and piloerection, and ten of twelve MCT rats died by *day 35*.

Electrocardiograms. ECG leads I and II were recorded every week after MCT injection. Figure 1 shows the representative tracings of ECG lead II recorded in

Table 1. *Body and organ weights in control and MCT-treated rats*

	n	BW, g	RV/BW, ×10⁻³	(LV + S)/BW, ×10⁻³	RV/(LV + S)	Lung/BW, ×10⁻³	Liver/BW, ×10⁻³	Kidney/BW, ×10⁻³
Day 0								
Control	6	144 ± 5.1	0.75 ± 0.04	2.88 ± 0.06	0.26 ± 0.02	7.00 ± 0.17	41.5 ± 4.23	9.70 ± 0.36
MCT	6	145 ± 6.2	0.77 ± 0.06	2.89 ± 0.06	0.27 ± 0.03	7.00 ± 0.55	42.9 ± 2.22	9.50 ± 0.28
Day 14								
Control	5	243 ± 4.2	0.60 ± 0.02	2.39 ± 0.12	0.26 ± 0.02	6.42 ± 0.36	44.8 ± 2.28	8.93 ± 0.18
MCT	5	233 ± 7.1	0.89 ± 0.06*	2.54 ± 0.11	0.35 ± 0.02*	8.26 ± 0.07*	46.5 ± 0.81	9.23 ± 0.57
Day 28								
Control	8	319 ± 8.0	0.64 ± 0.02	2.36 ± 0.12	0.27 ± 0.01	5.80 ± 0.57	42.3 ± 1.50	7.60 ± 0.35
MCT	13	251 ± 5.1*	1.45 ± 0.06*	2.51 ± 0.06	0.58 ± 0.03*	10.0 ± 0.55*	40.9 ± 2.00	7.60 ± 0.18

Values are means ± SE. MCT, monocrotaline; BW, body weight; RV, right ventricle; LV + S, left ventricle + septum. * Significantly different from control at $P < 0.05$.

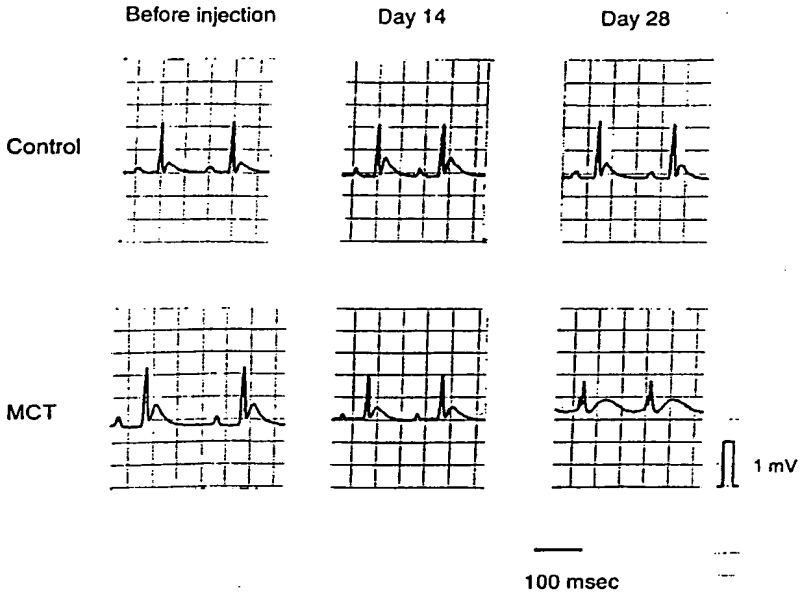


Fig. 1. Body surface electrocardiograms (ECG) of extremity lead (II) before monocrotaline (MCT) and saline injection and on *days 14* and *28* after injection. *Top traces* are records from a control rat, and *bottom traces* are from a MCT rat.

the same rats immediately before injection (*day 0*) and on *days 14* and *28* after the injection. In the ECGs of the MCT rat, the T wave was flattened and the Q-T interval was prolonged on *day 14* and the prolongation was remarkable on *day 28*, whereas the ECGs of the control rat remained unchanged over the entire observation period. Table 2 summarizes the ECG data obtained from control and MCT-treated rats on *days 0*, *14*, and *28* after injection. The Q-T interval was significantly prolonged, on average, by 9.5 and 10.2% on *days 14* and *28*, respectively, and the interval corrected by the heart rate ($Q\text{-}T_c$) was also significantly prolonged by 6.9 and 13.7% on *days 14* and *28*. Although QRS duration did not show any difference on *day 14*, it was prolonged by 16.2% on *day 28*. P-R interval was unchanged over the observation period.

Gene expression of Kv channels. To examine the effects of cardiac hypertrophy on mRNA expression of cloned K^+ channels, mRNA levels were measured with RPA, using the hearts obtained from control and MCT-treated rats killed on *day 28*. mRNA levels of three *Shaker* (Kv1.2, 1.4, 1.5), one *Sha*b (Kv2.1), and two *Sha*l (Kv4.2 and Kv4.3) channels were examined. These mRNAs were readily detected. Cyclophilin mRNA expression levels were used for the internal control.

Table 2. *ECG parameters recorded from control and MCT-treated rats*

	n	QRS Time, ms	P-R Time, ms	Q-T Interval, ms	$Q\text{-}T_c$
Day 0					
Control	5	24.0 ± 2.4	41.3 ± 0.5	96.6 ± 3.8	8.2 ± 0.2
MCT	5	24.2 ± 1.8	39.5 ± 9.1	96.2 ± 5.8	8.3 ± 0.3
Day 14					
Control	6	24.0 ± 2.4	39.0 ± 4.0	100.8 ± 3.6	8.7 ± 0.1
MCT	5	24.6 ± 1.8	43.0 ± 4.0	110.4 ± 1.9*	9.3 ± 0.1*
Day 28					
Control	6	26.4 ± 2.6	41.5 ± 4.9	98.8 ± 4.9	8.8 ± 0.5
MCT	5	31.5 ± 5.8*	39.0 ± 7.7	110.0 ± 4.2*	10.2 ± 0.4*

Values are means ± SE. $Q\text{-}T_c$, corrected Q-T interval. *Significantly different from control at $P < 0.05$.

Figure 2 shows the results for the *Shaker* family channels. The expression levels of Kv1.2 and Kv1.5 channels normalized to cyclophilin expression levels were significantly lower in the MCT-treated rats than in control rats [Kv1.2 (control vs. MCT): 1.0 ± 0.12 vs. 0.4 ± 0.13 (arbitrary units) ($n = 3$), $P < 0.05$; Kv1.5: 1.0 ± 0.05 vs. 0.24 ± 0.03 ($n = 3$), $P < 0.01$] (Fig. 2). Unlike Kv1.2 and Kv1.5, the expression levels of Kv1.4 mRNA did not show a significant difference between two groups [Kv1.4 (control vs. MCT): 1.0 ± 0.28 vs. 0.88 ± 0.44 ($n = 3$); not significant (NS)] (Fig. 2). The expression levels of Kv2.1 mRNA channels were significantly decreased in the MCT-treated rats compared with control rats [Kv2.1 (control vs. MCT): 1.0 ± 0.02 vs. 0.42 ± 0.05 ($n = 3$); $P < 0.05$] (Fig. 3). Figure 4 shows the expression levels of the *Sha*l family channels. The expression levels of Kv4.2 were markedly decreased in the MCT rats [Kv4.2 (control vs. MCT): 1.0 ± 0.08 vs. 0.19 ± 0.05 ($n = 3$); $P < 0.01$]. In the meantime, Kv4.3 mRNA expression levels were also significantly decreased, but the extent of decrease was moderate compared with that of Kv4.2 [Kv4.3 (control vs. MCT): 1.0 ± 0.05 vs. 0.55 ± 0.12 ($n = 3$); $P < 0.05$].

mRNA levels of Kv channels on *day 14* after injection were also measured to examine the effects of cardiac hypertrophy at the early stage. There was a slight increase in Kv1.4, Kv4.2, and Kv4.3 [Kv1.4 (control vs. MCT): 1.0 ± 0.22 vs. 1.27 ± 0.37 ($n = 3$), NS; Kv4.2: 1.0 ± 0.15 vs. 1.22 ± 0.28 ($n = 3$), NS; Kv4.3: 1.0 ± 0.20 vs. 1.21 ± 0.28 ($n = 3$), NS] and a slight decrease in Kv1.2 and Kv1.5 [Kv1.2 (control vs. MCT): 1.0 ± 0.25 vs. 0.87 ± 0.18 ($n = 3$), NS; Kv1.5: 1.0 ± 0.16 vs. 0.91 ± 0.28 ($n = 3$), NS], but these differences did not reach statistical significance. The mRNA level of Kv2.1 was

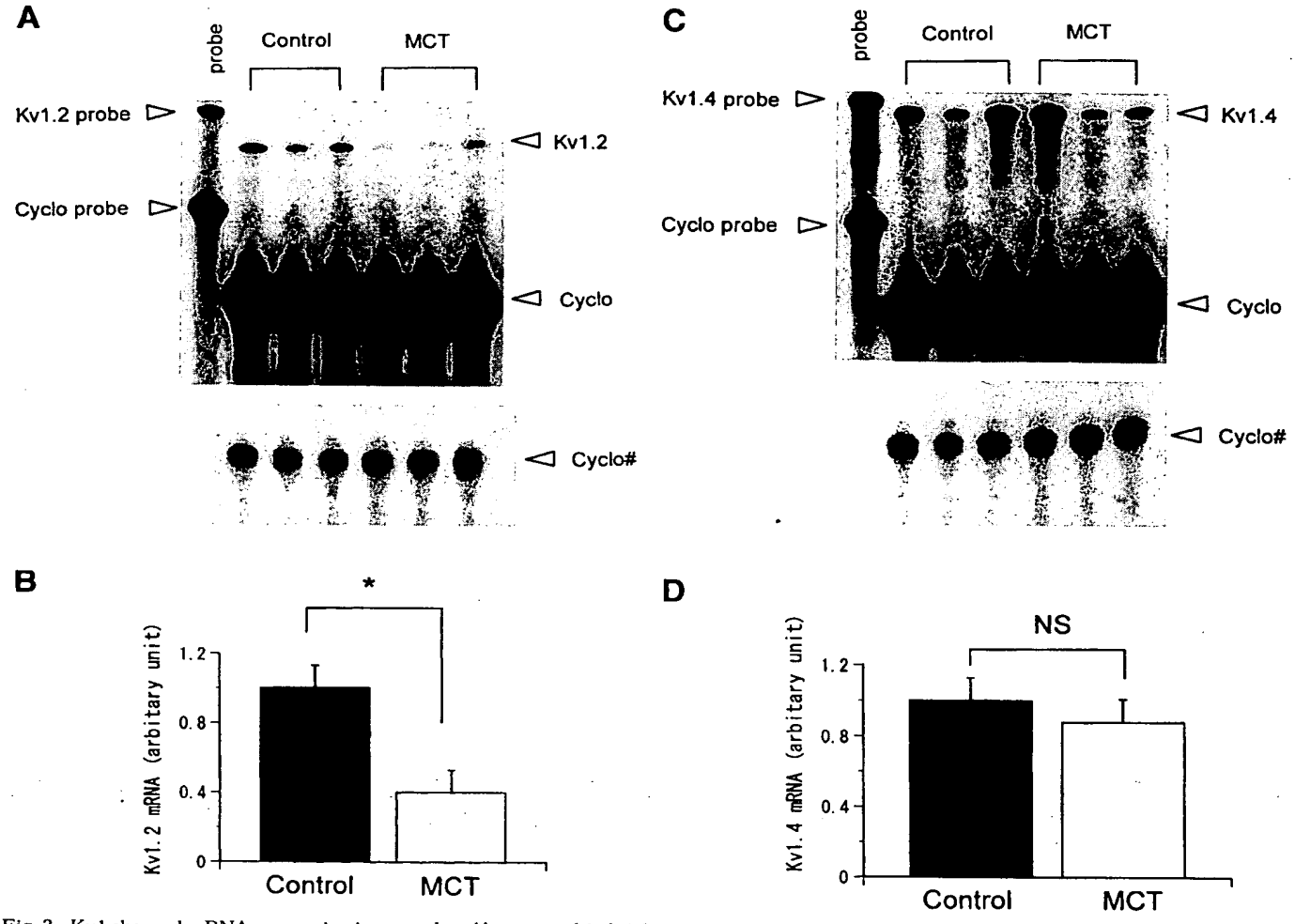


Fig. 2. Kv1 channel mRNA expression in control and hypertrophied right ventricles. Hypertrophied right ventricles were obtained from adult rats with pulmonary hypertension induced by a single injection of MCT. Total RNAs were extracted from right ventricles for ribonuclease protection assay on *day 28* after MCT injection. *A*, *C*, and *E*: mRNA expression of Kv1.2, Kv1.4, and Kv1.5, respectively. Cyclo, cyclophilin mRNA as internal control; Cyclo#, mRNA of cyclophilin with less exposure. *B*, *D*, and *F*: average amounts of Kv1.2 (*B*), Kv1.4 (*D*) and Kv1.5 (*F*) mRNA are presented as ratio to internal control ($n = 3$). Significantly different from control: * $P < 0.05$, ** $P < 0.01$. NS, not significant.

identical between groups [1.0 ± 0.26 vs. 0.99 ± 0.25 ($n = 3$), NS].

DISCUSSION

MCT caused hypertrophy in RV. In the present study, we investigated the underlying molecular mechanisms of the altered repolarization in ventricular hypertrophy, using rats with RV hypertrophy secondary to MCT-induced pulmonary hypertension. MCT is known to cause pulmonary hypertension in rats through endothelial cell damage, medial thickening of the muscular pulmonary arteries, and neomuscularization of nonmuscular distal arteries. A single injection of MCT caused macroscopic RV hypertrophy without any morphological changes in the LV. An increase of the ratios of RV and lung weights to BW was observed in the MCT-treated rats, but the ratios of kidney and liver weights to BW were not affected by the treatment. Recent experimental studies have indicated that an increase of endogenous endothelin-1, a potent endothelium-derived vasoconstrictor peptide, is involved in the pathogenesis of MCT-induced pulmonary hypertension (22, 25). However, the lack of morphological change in the LV may suggest that the hypertrophy may not be the result of direct action of this compound on the heart but the result of pressure overload caused by pulmonary hypertension. The MCT rats on *day 14* are considered to be in a compensated state of hypertrophy, because they showed normal growth and no physical signs of right-sided heart failure. On the other hand, the MCT rats on *day 28* had more of the properties of heart failure, because they showed a significant decrease in BW and physical signs of right-sided heart failure, including tachypnea, ascites, pleural effusion, edema-

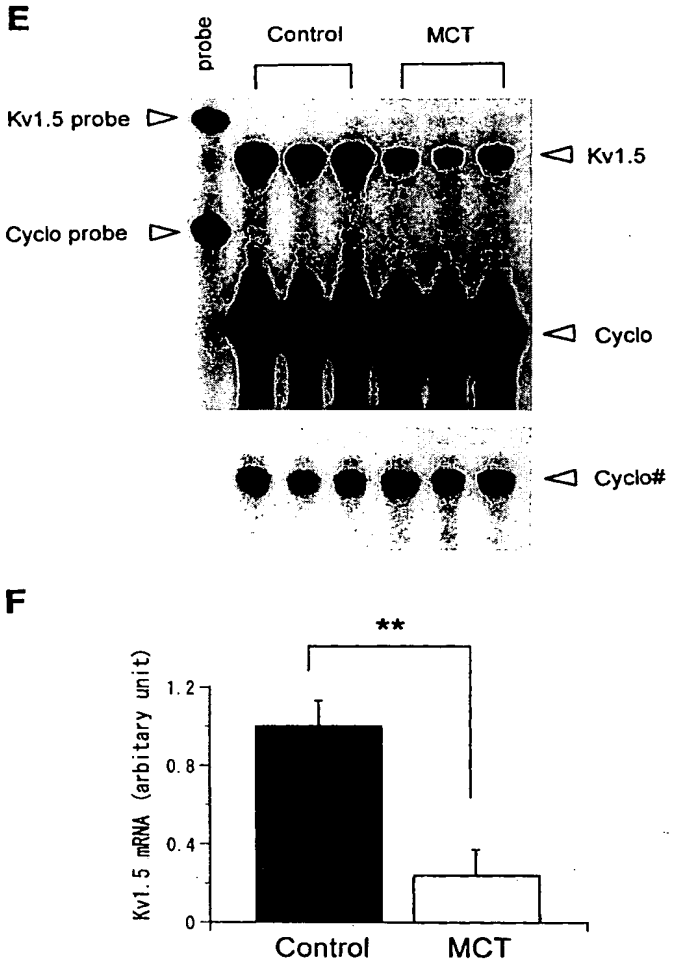


Fig. 2*E-F—Continued.*

tous extremities, and piloerection, in the following week.

Electrophysiological alterations in ventricular hypertrophy. In association with the development of hypertrophy, body surface electrocardiograms showed prolongation of Q-T and Q-T_c intervals and QRS duration, whereas P-R intervals were not affected (Table 2). In our previous electrophysiological experiments on single myocytes isolated from MCT-treated rats, cell membrane capacitance and APD of RV cells were increased progressively from *day 14* to *day 28*, whereas other parameters of the action potential (resting membrane potential and action potential amplitude) were unaffected (17). The APD at 90% repolarization of the MCT-treated RV cells was increased to 192% of control (n = 10) on *day 28* after the injection. As to the change of ionic currents responsible for the APD prolongation at the late stage of MCT-treated rats, we reported a significant reduction of I_{to} without any changes in its voltage dependence and inactivation kinetics (17). The changes in cell membrane capacitance and action potential configuration observed in our RV hypertrophy model are qualitatively similar to those in the reports by other investigators on the LV hypertrophy induced in rats by aortic banding and renovascular hypertension (4, 19, 32).

Molecular mechanisms of altered repolarization in ventricular hypertrophy. We have shown in the present study that mRNA levels of Kv4.2 and Kv4.3 are decreased in MCT-treated rats at the advanced stage of hypertrophy. Kv4.2 and Kv4.3 are supposed to be the most likely candidates for I_{to} in adult rat ventricles (3, 9). The reduction of I_{to} density at the advanced stage of RV hypertrophy in MCT-treated rats could be the result of downregulation of Kv4.2 and Kv4.3 gene expression. In experiments using renovascular hypertensive rats, Takimoto et al. (30) demonstrated significant reduction of mRNA levels for Kv4.2 and Kv4.3 in the LV in association with the progress of hypertrophy but no significant changes in mRNA levels for *Shaker*-related

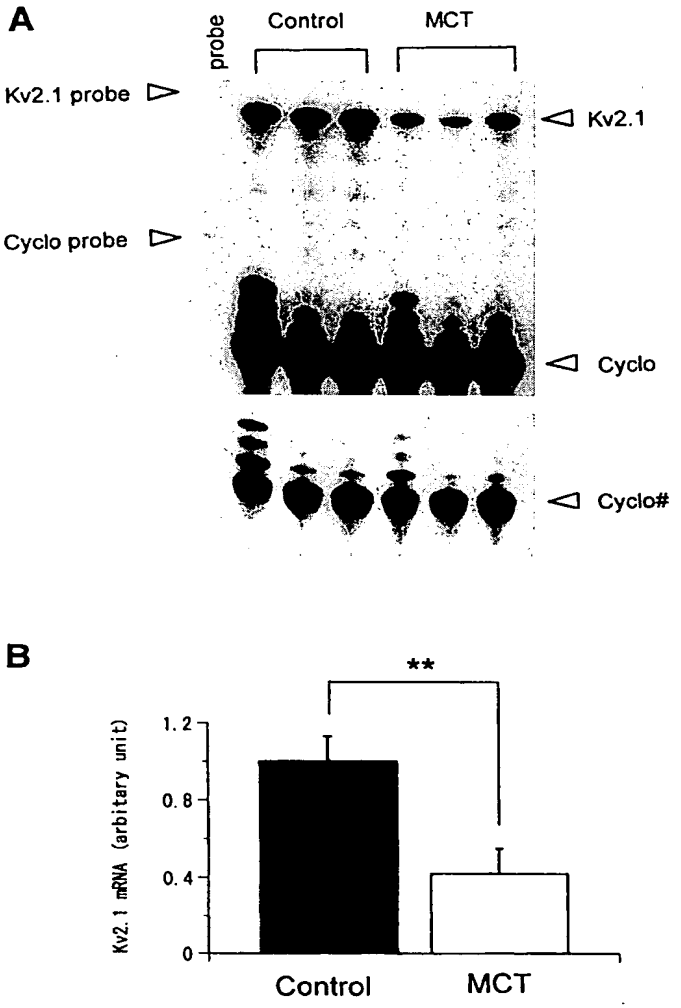


Fig. 3. Kv2.1 mRNA levels in right ventricles of control and MCT-treated rats on *day 28* after injection. *A*: Kv2.1 mRNA measured by ribonuclease protection assay. Cyclo, cyclophilin mRNA as internal control; Cyclo*, mRNA of cyclophilin with less exposure. *B*: average amount of Kv2.1 mRNA is presented as ratio to internal control (n = 3). ** Significantly different from control at $P < 0.01$.

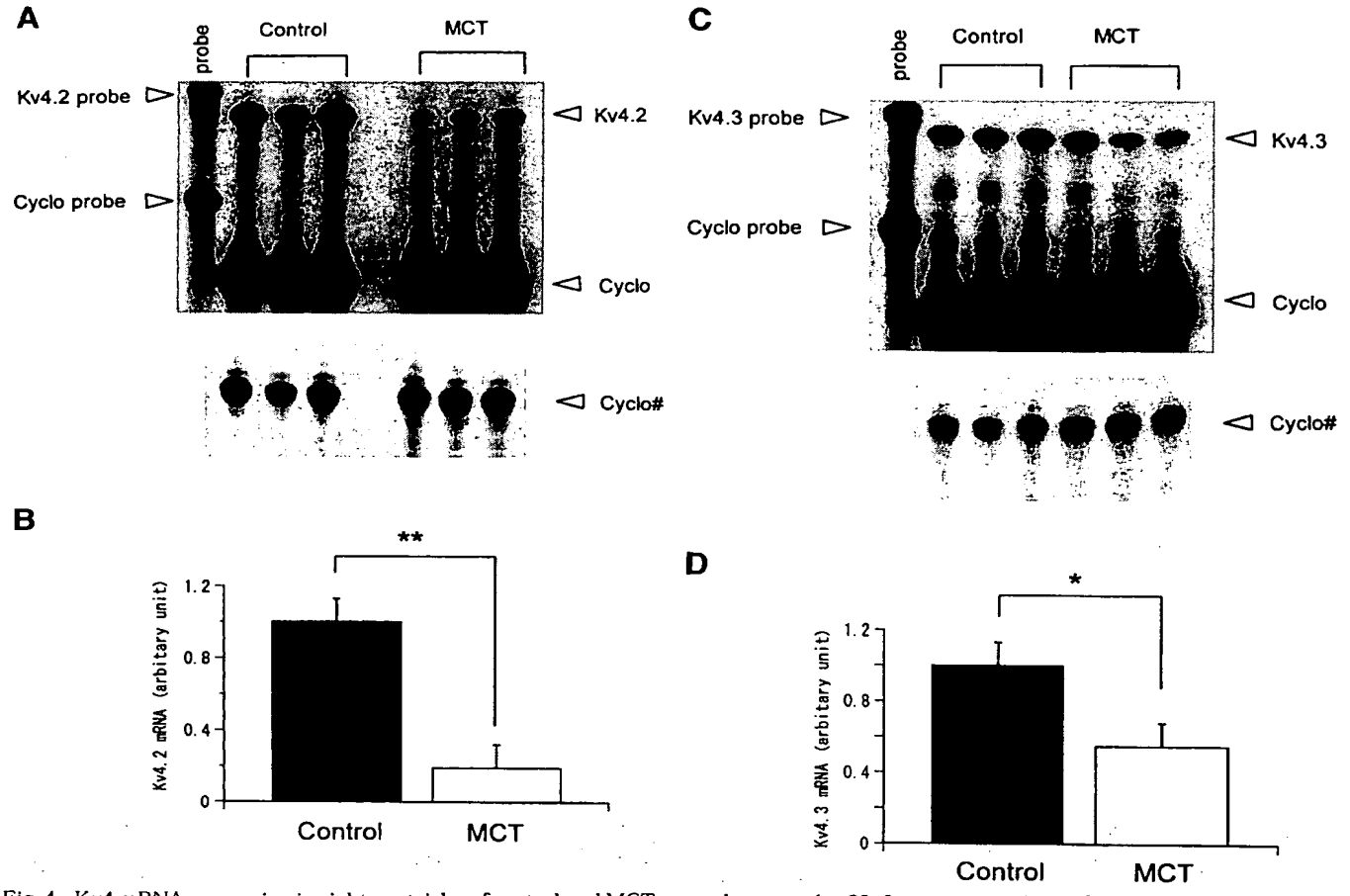


Fig. 4. Kv4 mRNA expression in right ventricles of control and MCT-treated rats on *day 28* after injection. *A* and *C*: Kv4.2 (*A*) and Kv4.3 (*C*) mRNA were measured by ribonuclease protection assay. Cyclo, cyclophilin mRNA as internal control; Cyclo#, mRNA of cyclophilin with less exposure. *B* and *D*: average amounts of Kv4.2 (*B*) and Kv4.3 (*D*) are presented as ratio to internal control ($n = 3$). Significantly different from control: * $P < 0.05$, ** $P < 0.01$.

(Kv1.2, Kv1.4, Kv1.5), *Sha*b-related (Kv2.1), and KvLQT1 channels.

In our experiments using MCT-treated rats, the mRNA levels for Kv1.2, Kv1.5, and Kv2.1 were also decreased significantly at the advanced stage of RV hypertrophy. Reasons for the discrepancy between our data and those reported by Takimoto et al. (30) are unclear; it might be related to different procedures used to produce hypertrophy.

The physiological and pathological roles of these cloned voltage-gated K^+ channels in native cardiac cells are still unsettled (4, 8). Heterologous expression of Kv1.2, Kv1.5, and Kv2.1 in *Xenopus* oocytes has been shown to cause delayed-rectifier type current (I_K) or rapidly activating sustained outward currents (I_{sus}, I_{ss}, or I_{Kur}), which are sensitive to 4-AP and TEA. In adult rat ventricular cells, the amplitude of these delayed-rectifier or sustained type outward currents is much less than that of I_{to}. This makes it difficult to detect the change of their current density in association with the progress of ventricular hypertrophy. Nevertheless, we cannot rule out some obligatory roles of the downregulation of Kv1.2, Kv1.5, and Kv2.1 gene expression in the APD prolongation in hypertrophied ventricular cells.

As for the early stage of hypertrophy, there was no significant difference in mRNA levels of these Kv channels. Thus there is a discrepancy between the increase of I_{to} density and the mRNA levels of the channels. At present, there is no clear interpretation for this discrepancy. There might be unknown subcellular factors that affect the protein synthesis or the availability of the channels in the pathological condition.

Limitation of study. Because the presence of mRNA does not necessarily mean the presence of the encoded proteins, studies measuring the mRNA levels have limitations for the understanding of the mechanism of the pathophysiological changes. To further elucidate the mechanism, studies measuring protein levels, including Western blot analysis and immunohistochemistry, will be required.

Address for reprint requests and other correspondence: I. Kodama, Dept. of Circulation, Research Institute of Environmental Medicine, Nagoya Univ., Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan (E-mail: ikodama@riem.nagoya-u.ac.jp).

Received 10 September 1998; accepted in final form 10 June 1999.

REFERENCES

1. **Aronson, R. S.** Characteristics of action potentials of hypertrophied myocardium from rats with renal hypertension. *Circ. Res.* 47: 443–454, 1980.
2. **Baldwin, T. J., M. L. Tsaur, G. A. Lopez, Y. N. Jan, and L. Y. Jan.** Characterization of a mammalian cDNA for an inactivating voltage-sensitive K^+ channel. *Neuron* 7: 471–483, 1991.
3. **Barry, D. M., and J. M. Nerbonne.** Differential expression of voltage-gated K^+ channel subunits in adult rat heart. Relation to functional K^+ channels? *Circ. Res.* 77: 361–369, 1995.
4. **Barry, D. M., and J. M. Nerbonne.** Myocardial potassium channels: electrophysiological and molecular diversity. *Annu. Rev. Physiol.* 58: 363–394, 1996.
5. **Bénitah, J.-P., A. M. Gomezu, P. Bailly, J.-P. Da Ponte, G. Berson, C. Delgado, and P. Lorente.** Heterogeneity of the early outward current in ventricular cells isolated from normal and hypertrophied rat heart. *J. Physiol. (Lond.)* 469: 111–138, 1993.
6. **Beuckelmann, D. J., M. Nabauer, and E. Erdmann.** Alteration of K^+ currents in isolated human ventricular myocytes from patients with terminal heart failure. *Circ. Res.* 73: 379–385, 1993.
7. **Cerbai, E., M. Barbieri, Q. Li, and A. Mugelli.** Ionic basis of action potential prolongation of hypertrophied myocytes isolated from hearts of spontaneously hypertensive rats of different ages. *Cardiovasc. Res.* 28: 1180–1187, 1994.
8. **Deal, K. K., S. K. England, and M. M. Tamkun.** Molecular physiology of cardiac potassium channels. *Physiol. Rev.* 76: 49–67, 1996.
9. **Dixon, J. E., and D. McKinnon.** Quantitative analysis of potassium channel mRNA expression in atrial and ventricular muscle of rats. *Circ. Res.* 75: 252–260, 1994.
10. **Dixon, J. E., W. Shi, H.-S. Wang, C. McDonald, H. Yu, R. S. Wymore, I. S. Cohen, and D. McKinnon.** Role of the Kv4.3 K^+ channel in ventricular muscle. A molecular correlate for the transient outward current. *Circ. Res.* 79: 659–668, 1996.
11. **Frech, G. C., A. M. VanDongen, G. Schuster, A. M. Brown, and R. H. Joho.** A novel potassium channel with delayed rectifier properties isolated from rat brain by expression cloning. *Nature* 340: 642–645, 1989.
12. **Furukawa, T., R. J. Myerburg, N. Furukawa, S. Kimura, and A. L. Bassett.** Metabolic inhibition of $I_{Ca,L}$ and I_K differs in feline left ventricular hypertrophy. *Am. J. Physiol.* 266 (*Heart Circ. Physiol.* 35): H1121–H1131, 1994.
14. **Hayashi, Y., J. F. Hussa, and J. Lalich.** Cor pulmonale in rats. *Lab. Invest.* 16: 875–881, 1967.
15. **Keung, E. C.** Calcium current is increased in isolated adult myocytes from hypertrophied rat myocardium. *Circ. Res.* 64: 753–763, 1989.
16. **Kleiman, R. B., and S. R. Houser.** Calcium currents in normal and hypertrophied isolated feline ventricular myocytes. *Am. J. Physiol.* 255 (*Heart Circ. Physiol.* 24): H1434–H1442, 1988.
17. **Lee, J. K., I. Kodama, H. Honjo, T. Anno, K. Kamiya, and J. Toyama.** Stage-dependent changes in membrane currents in rats with monocrotaline-induced right ventricular hypertrophy. *Am. J. Physiol.* 272 (*Heart Circ. Physiol.* 41): H2833–H2842, 1997.
18. **Levy, D., R. J. Garrison, D. D. Savage, W. B. Kannel, and W. P. Castelli.** Prognostic implications of echocardiography determined left ventricular mass in the Framingham heart study. *N. Engl. J. Med.* 322: 1561–1566, 1990.
19. **Li, Q., and E. C. Keung.** Effects of myocardial hypertrophy on transient outward current. *Am. J. Physiol.* 266 (*Heart Circ. Physiol.* 35): H1738–H1745, 1994.
20. **Matsubara, H., J. Suzuki, and M. Inada.** Shaker-related potassium channel, Kv1.4, mRNA regulation in cultured rat heart myocytes and differential expression of Kv1.4 and Kv1.5 genes in myocardial development and hypertrophy. *J. Clin. Invest.* 92: 1659–1666, 1993.
21. **McKinnon, D.** Isolation of a cDNA clone coding for putative second potassium channel indicates the existence of a gene family. *J. Biol. Chem.* 264: 8230–8236, 1989.
22. **Miyauchi, T., R. Yorikane, S. Sakai, T. Sakurai, M. Okada, M. Nishikibe, M. Yano, I. Yamaguchi, Y. Sugita, and K. Goto.** Contribution of endogenous endothelin-1 to the progression of cardiopulmonary alterations in rats with monocrotaline-induced pulmonary hypertension. *Circ. Res.* 73: 887–897, 1993.
23. **Nishiyama, A., F. Kambe, K. Kamiya, S. Yamaguchi, Y. Murata, H. Seo, and J. Toyama.** Effects of thyroid and glucocorticoid hormones on Kv1.5 potassium channel gene expression in the rat left ventricle. *Biochem. Biophys. Res. Commun.* 237: 521–526, 1997.
24. **Nuss, H. B., and S. R. Houser.** Voltage dependence of contraction and calcium current in severely hypertrophied feline ventricular myocytes. *J. Mol. Cell. Cardiol.* 23: 717–726, 1991.
25. **Okada, M., C. Yamashita, M. Okada, and K. Okada.** Role of endothelin-1 in beagles with dehydromonocrotaline-induced pulmonary hypertension. *Circ. Res.* 92: 114–119, 1995.
26. **Ryder, K. O., S. M. Bryant, and G. Hart.** Membrane current changes in left ventricular myocytes isolated from guinea pigs after abdominal aortic coarctation. *Cardiovasc. Res.* 27: 1278–1287, 1993.
27. **Scamps, F., E. Mayoux, D. Charlemagne, and G. Vassort.** Calcium current in single cells isolated from normal and hypertrophied rat heart. Effects of β-adrenergic stimulation. *Circ. Res.* 67: 199–208, 1990.
28. **Swanson, R., J. Marshall, and J. S. Smith.** Cloning and expression of cDNA and genomic clones encoding three delayed rectifier potassium channels in rat brain. *Neuron* 4: 929–939, 1990.
29. **Takimoto, K., and E. S. Levitan.** Glucocorticoid induction of Kv1.5 K^+ channel gene expression in ventricle of rat heart. *Circ. Res.* 75: 1006–1013, 1994.
30. **Takimoto, K., D. Li, K. M. Hershman, P. Li, E. K. Jackson, and E. S. Levitan.** Decreased expression of Kv4.2 and novel Kv4.3 K^+ channel subunit mRNAs in ventricles of renovascular hypertensive rats. *Circ. Res.* 81: 533–539, 1997.
31. **Tomita, F., A. L. Bassett, R. J. Myerburg, and S. Kimura.** Diminished transient outward currents in rat hypertrophied ventricular myocytes. *Circulation* 75: 296–303, 1994.
32. **Tseng, C. J., G. N. Tseng, A. Schwartz, and M. A. Tanouye.** Molecular cloning and functional expression of a potassium channel cDNA isolated from a rat cardiac library. *FEBS Lett.* 268: 63–68, 1990.

1: J Cardiovasc Electrophysiol. 2000 Nov;11(11):1252-61.

Early down-regulation of K+ channel genes and currents in the postinfarction heart.

Huang B, Qin D, El-Sherif N.

Department of Medicine, State University of New York Health Science Center, Brooklyn 11203, USA.

INTRODUCTION: Down-regulation of key K+ channel subunit gene expression and K+ currents is a universal response to cardiac hypertrophy, whatever the cause, including the postmyocardial infarction (post-MI) remodeled heart. METHODS AND RESULTS: We investigated the hypothesis that down-regulation of K+ channel genes and currents post-MI occurs early and before significant remodeled hypertrophy of the noninfarcted myocardium could be detected. We investigated (1) the incidence of induced ventricular tachyarrhythmias (VT) in 3-day post-MI rat heart; (2) action potential (AP) characteristics of isolated left ventricular (LV) myocytes from sham-operated and 3-day post-MI heart; (3) time course of changes in outward K+ currents Ito-fast(f) and I(K) in isolated myocytes from 3-day and 4-week post-MI noninfarcted LV and compared the changes with sham-operated animals; and (4) changes in the messenger and protein levels of Kv2.1, Kv4.2, and Kv4.3 in the LV and right ventricle of 3-day post-MI heart. Sustained VT was induced in 6 of 10 3-day post-MI rats and in none of 8 sham rats. The membrane capacitance of myocytes isolated from 3-day post-MI noninfarcted LV was not significantly different from control, whereas membrane capacitance 4-week post-MI was significantly higher, reflecting the development of hypertrophy. AP duration was increased and the density of Ito-f and I(K) were significantly decreased in 3-day post-MI LV myocytes compared with sham. The reduced density of Ito did not significantly differ in 4-week post-MI LV myocytes, whereas the density of I(K) was decreased further at 4 weeks post-MI. The changes in Ito-f and I(K) correlated with decreased messenger and protein levels of Kv4.2/Kv4.3 and Kv2.1, respectively. CONCLUSION: These results support the hypothesis that down-regulation of K+ channel gene expression and current in the post-MI LV occurs early and may be dissociated from the slower time course of post-MI remodeled hypertrophy. These changes may contribute to early arrhythmogenesis of the post-MI heart.

PMID: 11083246 [PubMed - indexed for MEDLINE]

Display Abstract Show: 20 Sort Send to Text

Feb 25 2005 07:07:33

0888-8809/01/$03.00/0
Printed in U.S.A.

EXHIBIT J

Molecular Endocrinology 15(8):1423–1435

Members of the Kv1 and Kv2 Voltage-Dependent K^+ Channel Families Regulate Insulin Secretion

PATRICK E. MACDONALD, XIAO FANG HA, JING WANG, SIMON R. SMUKLER, ANTHONY M. SUN, HERBERT Y. GAISANO, ANN MARIE F. SALAPATEK, PETER H. BACKX, AND MICHAEL B. WHEELER

Departments of Medicine (H.Y.G., P.H.B., M.B.W.) and Physiology (P.E.M., X.F.H., J.W., S.R.S., A.M.S., A.M.F.S., P.H.B., M.B.W.), University of Toronto, Toronto Ontario, Canada M5S 1A8

In pancreatic β-cells, voltage-dependent K^+ (Kv) channels are potential mediators of repolarization, closure of Ca^{2+} channels, and limitation of insulin secretion. The specific Kv channels expressed in β-cells and their contribution to the delayed rectifier current and regulation of insulin secretion in these cells are unclear. High-level protein expression and mRNA transcripts for Kv1.4, 1.6, and 2.1 were detected in rat islets and insulinoma cells. Inhibition of these channels with tetraethylammonium decreased I_{DR} by approximately 85% and enhanced glucose-stimulated insulin secretion by 2- to 4-fold. Adenovirus-mediated expression of a C-terminal truncated Kv2.1 subunit, specifically eliminating Kv2 family currents, reduced delayed rectifier currents in these cells by 60–70% and enhanced glucose-stimulated insulin secretion from rat islets by 60%. Expression of a C-terminal truncated Kv1.4 subunit, abolishing Kv1 channel family currents, reduced delayed rectifier currents by approximately 25% and enhanced glucose-stimulated insulin secretion from rat islets by 40%. This study establishes that Kv2 and 1 channel homologs mediate the majority of repolarizing delayed rectifier current in rat β-cells and that antagonism of Kv2.1 may prove to be a novel glucose-dependent therapeutic treatment for type 2 diabetes. (*Molecular Endocrinology* 15: 1423–1435, 2001)

THE ABILITY OF pancreatic islets of Langerhans to secrete insulin in response to increased blood glucose levels is essential for the maintenance of normoglycemia. Dysregulation of islet insulin secretion is at least partly responsible for the development of type 2 diabetes mellitus (1). In the β-cell, glucose stimulation is coupled to insulin secretion through voltage-dependent and voltage-independent mechanisms (2, 3). Voltage-dependent mechanisms of stimulus-secretion coupling are better characterized and are described in a number of reviews (4–7). Briefly, increased glucose metabolism in pancreatic β-cells, resulting from high postprandial blood glucose, increases the intracellular ATP:ADP ratio. This leads to closure of ATP-sensitive K^+ (K_{ATP}) channels and depolarization of the cell membrane (8), an effect mimicked by the sulfonylurea drugs independent of blood glucose (9, 10).

Depolarization of the β-cell membrane results in the opening of L-type Ca^{2+} channels, increasing the intracellular Ca^{2+} concentration ($[Ca^{2+}]_i$) and ultimately stimulating insulin secretion. β-Cell repolarization is mediated by a delayed rectifier current (I_{DR}) similar to those generated by voltage-dependent K^+ (Kv) or Ca^{2+}-sensitive voltage-dependent K^+ (K_{Ca}) channels (5, 11–14). Accordingly, overexpression of a Kv channel in transgenic mice was associated with hyperglycemia and hypoinsulinemia, and in an insulinoma cell line this manipulation attenuated $[Ca^{2+}]_i$ increases associated with glucose stimulation (15). In addition, inhibitors of I_{DR} are known to enhance $[Ca^{2+}]_i$ oscillations (16) and insulin secretion (11, 13) in a glucose-dependent manner.

There are at least 11 currently known Kv channel families containing 26 homologs (17–22), and of these, members of the Kv1, Kv2, and Kv3 channel families mediate currents similar to those observed in pancreatic β-cells (5, 23–25). The task of identifying the channel homologs responsible for repolarization of pancreatic β-cells is difficult because heterotetrameric Kv channels and channels associated with regulatory β-subunits often do not exhibit the electrical and pharmacological properties of the constituent pore-forming subunits (17, 26–29).

Despite previous studies showing that insulin-secreting cells express mRNA transcripts for a number of Kv and K_{Ca} channels (5) and Kv2.1 protein (11), no functional data exist for a role for specific channels or channel families in β-cell repolarization and the regulation of insulin secretion. We have now characterized the mRNA and protein expression of Kv1 and Kv2 channel family homologs in rat islets and insulinoma cell lines. Pharmacological agents and dominant-negative C-terminal truncated Kv1 (Kv1.4N) and Kv2 (Kv2.1N) channel subunit mutants were used to deter-

Abbreviations: $[Ca^{2+}]_i$, intracellular Ca^{2+} concentration; EGFP, enhanced green fluorescent protein; FT, freeze-thaw media; GSIS, glucose-stimulated insulin secretion; IBMX, 3-isobutyl-1-methylxanthine; HG-RPMI, high-glucose Roswell Park Memorial Institute medium; I_{DR}, delayed rectifier current; K_{Ca}, Ca^{2+}-sensitive voltage-dependent K^+ channel; KRB, Krebs Ringer bicarbonate; Kv, voltage-dependent K^+ channel; LG-RPMI, low-glucose RPMI; TEA, tetraethylammonium.

mine the role of specific channels in mediating I_{DR} and regulating insulin secretion in the glucose-responsive HIT-T15 cell line and in rat islets.

RESULTS

Effect of I_{DR} Inhibition on Insulin Secretion

HIT-T15 cells or rat islets were incubated with the general Kv and K_{Ca} channel antagonist tetraethylammonium (TEA) at concentrations known to inhibit delayed rectifier currents while having minimal effects on K_{ATP} channels (12, 30, 31). In HIT-T15 cells, TEA (20 mM) enhanced glucose-stimulated insulin secretion (GSIS) (from 0.51 ± 0.10 to 1.43 ± 0.14 ng/ml/2 h, n = 15; $P < 0.001$), but most importantly, had no effect in the absence of glucose (Fig. 1A). Similarly, GSIS from rat islets was enhanced by TEA (20 mM) (from 0.17 ± 0.03, n = 24 to 0.81 ± 0.18 ng/islet/h, n = 13; $P < 0.01$), which had no effect in the absence of stimulatory concentrations of glucose (control, n = 23; 20 mM TEA, n = 13) (Fig. 1B). TEA enhanced GSIS from rat islets in a dose-dependent manner (Fig. 1C) with an EC_{50} of 8.24 mM (n = 9). The effects of TEA were not related to cellular toxicity, since a 2-h exposure (20 mM) did not affect the survival of HIT-T15 cells, as detected by propidium iodide fluorescence (not shown).

To determine whether TEA's insulinotropic activity was dependent upon depolarization through K_{ATP} channel closure and not glucose *per se*, we examined whether TEA could enhance insulin secretion stimulated by K_{ATP} channel inhibition in the absence of glucose. Micromolar concentrations of the K_{ATP} channel antagonist glyburide (Sigma, St. Louis, MO) have been shown to stimulate insulin secretion from HIT-T15 cells in the absence of glucose (32, 33). Glyburide (2 μM) simulated insulin secretion nearly 2-fold from HIT-T15 cells (from 0.14 ± 0.01, n = 8 to 0.25 ± 0.01 ng/ml/2 h, n = 8; $P < 0.001$) in the absence of glucose. Addition of TEA (20 mM) in the presence of glyburide enhanced insulin secretion further (to 0.56 ± 0.06 ng/ml/2 h, n = 8; $P < 0.01$ compared with glyburide alone) (Fig. 2).

Similarly, islets were incubated in nonstimulatory concentrations of glucose (2.5 mM) with TEA (20 mM) and/or the sulfonylurea drug glyburide. Glyburide at 2 μM elicited a large insulin response that was not enhanced by 20 mM TEA (Fig. 3A). Because the micromolar concentrations of glyburide commonly used to stimulate insulin secretion are approximately 4,000 times the published EC_{50} in rodent islets (34), nonspecific effects on ion channels or non-β-cells are possible. With 10 nM glyburide, TEA (20 mM) significantly enhanced insulin secretion compared with glyburide alone in both the presence (n = 10; $P < 0.05$) and absence (n = 12; $P < 0.05$) of stimulatory glucose (Fig. 3B). PKA pathway signaling enhances GSIS, partly

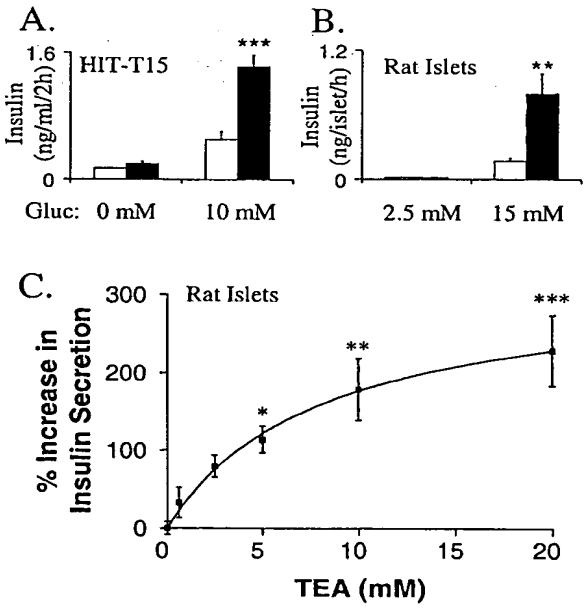


Fig. 1. I_{DR} Inhibition Enhances GSIS

The general Kv channel antagonist TEA (20 mM; *black bars*) enhanced insulin secretion from HIT-T15 insulinoma cells (A) and isolated rat islets (B) over 2 h compared to controls (*white bars*). This effect occurred only in the presence of stimulatory glucose. In rat islets, TEA dose-dependently enhanced insulin secretion stimulated by 15 mM glucose in a dose-dependent manner (C). The half-maximal effect of TEA was observed at 8.24 mM. *, $P < 0.05$, **, $P < 0.01$; and ***, $P < 0.001$ compared with controls.

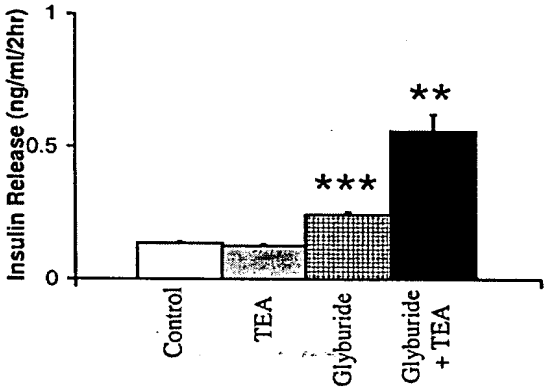


Fig. 2. I_{DR} Inhibition Enhances Glyburide-Stimulated Insulin Secretion from HIT-T15 Cells

In the absence of glucose, the K_{ATP} channel antagonist glyburide (2 μM; *hatched bar*) depolarizes HIT-T15 cells and stimulates insulin secretion compared with control (*white bar*). The general Kv channel antagonist TEA (20 mM) alone (*gray bar*) had no effect on unstimulated insulin secretion but further enhanced insulin secretion from HIT-T15 cells depolarized by glyburide (*black bar*). **, $P < 0.01$ and ***, $P < 0.001$ compared with control.

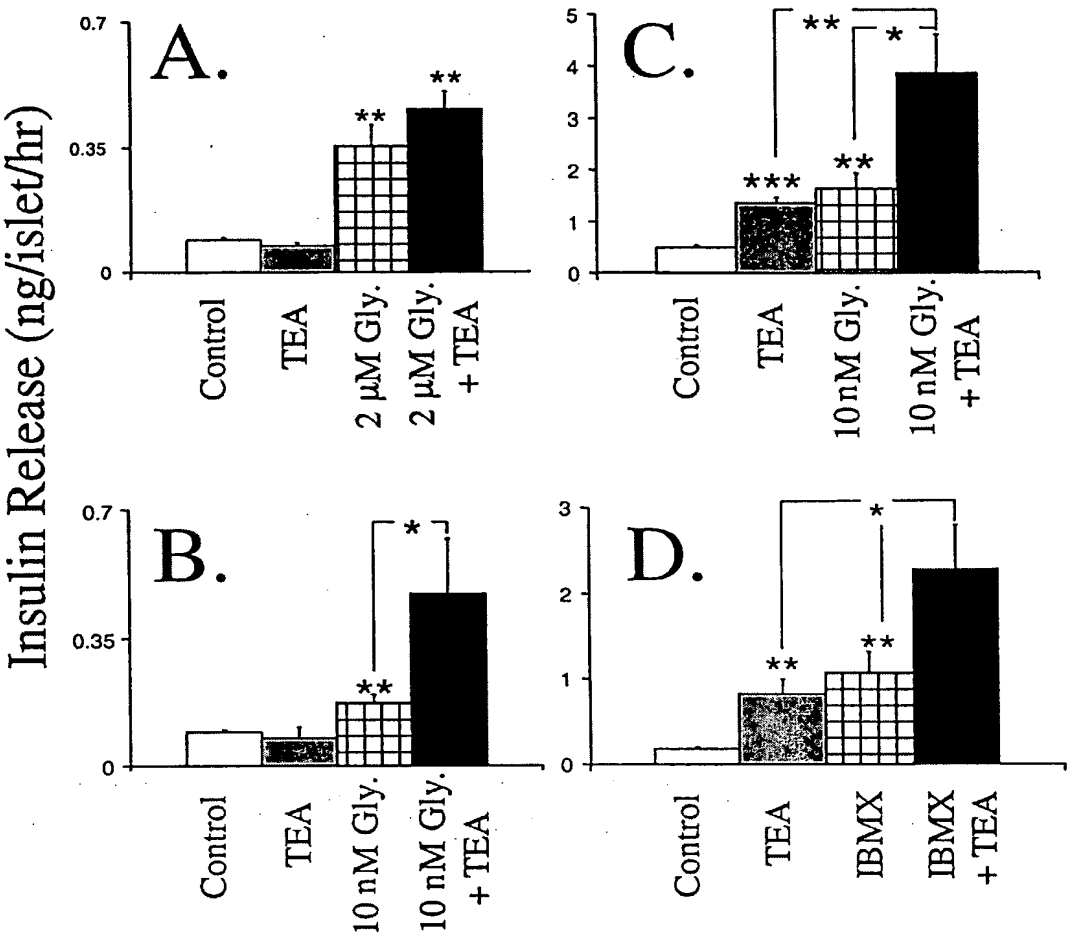


Fig. 3. I_{DR} Inhibition Enhances the Insulinotropic Effect of K_{ATP} and PKA Pathway Agonists

In the presence of 2.5 mM glucose (A and B) the K_{ATP} channel antagonist glyburide [*hatched bars*, at 2 μM (A) or 10 nM (B)] stimulates insulin secretion from isolated rat islets compared with controls (*white bars*). The general Kv channel antagonist TEA (20 mM; *gray bars*) had no effect on unstimulated insulin secretion, but further enhanced insulin secretion from isolated rat islets together (*black bars*) with 10 nM glyburide (B). With stimulatory glucose (15 mM, panels C and D), TEA (20 mM) enhanced insulin secretion and the effects of secretagogues acting through the K_{ATP} (panel C, 10 nM glyburide) and PKA (panel D, 1 μM IBMX) pathways. *, $P < 0.05$; **, $P < 0.01$; and ***, $P < 0.001$ compared with controls unless otherwise indicated.

through actions on ion channels (35, 36). In the present study, TEA (20 mM) enhanced the insulinotropic effect of the PKA pathway agonist 3-isobutyl-1-methylxanthine (IBMX) (1 μM) in the presence of stimulatory glucose (Fig. 3D). These results demonstrate that membrane depolarization is sufficient to allow TEA's insulinotropic effect and that TEA enhances the insulinotropic effects of agonists acting through the K_{ATP} and PKA pathways.

Pancreatic Islet and β-cell Kv Channel Expression

The above results demonstrate that the blockade of I_{DR} can enhance insulin secretion when glucose or channel antagonists close K_{ATP} channels. To determine which K^+ channels mediate I_{DR} in insulin-secreting cells, HIT-T15 cell and rat islet total RNA were examined for Kv gene transcripts via RT-PCR (Table 1). RT-PCR of HIT-T15 cell RNA with Kv1.1, 1.3, and 1.4 specific primers resulted in amplification products of the expected size. RT-PCR of rat islet RNA yielded cDNA fragments of the correct size for Kv1.2, 1.3, 1.4, 1.6, and 2.1. Sequencing confirmed that each fragment corresponded to the appropriate channel with a high degree of nucleotide and predicted amino acid identity with the respective human channel. All primer sets produced PCR products of the appropriate size upon RT-PCR of rat brain or mouse skeletal muscle (Kv1.7) total RNA as a positive control. No PCR product was visible in the water blank controls.

Table 1. Identification of Kv Channel Homologs in HIT-T15 Cells and Rat Islets by RT-PCR of Total RNA

Kv Transcript	Expected Product Size (bp)	HIT-T15 Cells		Rat Islets	
		% Sequence Identity[a]	% Amino Acid Identity[a]	% Sequence Identity[a]	% Amino Acid Identity[a]
1.1	270	92	93		
1.2	412			91	99
1.3	438	90	94	89	95
1.4	712	93	98	91	98
1.5	448				
1.6	735			84	89
1.7	580				
2.1	401			88	98
2.2	716				

[a] Identity as compared to human cDNA and protein sequences.

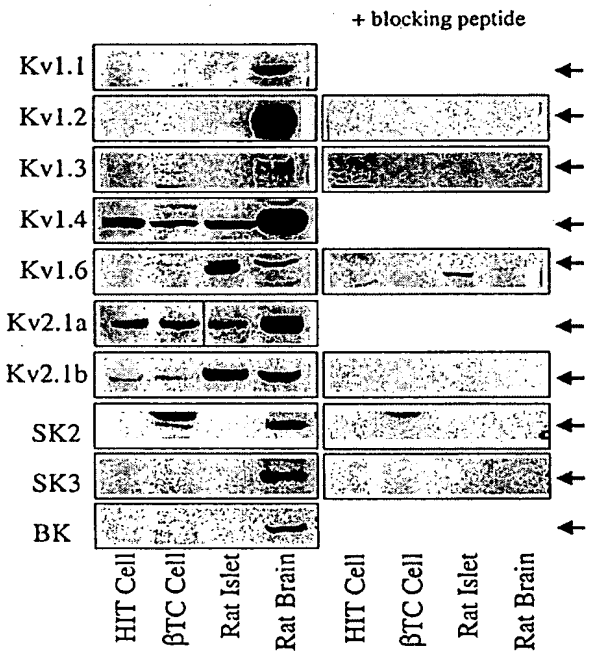


Fig. 4. Kv and K_{Ca} Channel Protein Expression

HIT-T15 cell, βTC-6f7 cell, rat islet, and rat brain lysates (50 μg protein) were probed for Kv1, Kv2, and K_{Ca} channel proteins using specific antibodies (see *Materials and Methods*). When available, control antigen (blocking peptide) was incubated with the channel antibody before probing of membranes to demonstrate the specificity of detection. Kv2.1 protein was detected with two separate antibodies (anti-Kv2.1a and anti-Kv2.1b). Anti-Kv2.1b was found to be more species specific for rat.

Western blot studies confirmed the protein expression of Kv1.4, 1.6, 2.1, and 1.2 (at lower levels) in rat islets (Fig. 4). Expression of Kv1.4 and 2.1 protein was detected in the HIT-T15 and βTC-6f7 insulinoma cell lines. Despite failure to detect Kv2.1 mRNA in HIT-T15 cells by RT-PCR, protein expression by this cell line is clearly abundant. It is possible that species selectivity of our primers resulted in our inability to detect the mRNA transcript in this hamster cell line. Levels of Kv2.1 protein detected in islets were roughly equivalent to those in the rat brain control using the Kv2.1b antibody. However, the levels of Kv2.1 detected differed markedly between the two antibodies, possibly reflecting variable species affinity (Fig. 4). Kv1.1 protein was detectable at low levels in HIT-T15 cells with longer exposures (not shown) but was not detected in rat islets. Specific protein bands for the K_{Ca} channels BK, SK2, and SK3 were not detected in insulin-secreting cells, with the exception of a light but detectable band for SK2 in βTC-6f7 cells (Fig. 4). Detection of Kv2.1 was used as a positive control in all protein samples from islets and β-cell lines.

Characterization of TEA-Sensitive I_{DR} in Insulin-Secreting Cells

As illustrated in Fig. 5, I_{DR} recorded from HIT-T15 cells or rat islet cells were noninactivating over 500 msec. Despite similar kinetic properties, current amplitudes (at the end of a 500-msec pulse to 70 mV from a holding potential of −70 mV) in HIT-T15 cells were approximately double those observed in rat islet cells (Fig. 5). Because of the inclusion of 1 mM EGTA and 5 mM MgATP within the pipette solution, the outward currents are expected to primarily reflect the opening of Kv channels, with minimal contributions from K_{Ca} or K_{ATP} channels. Since native β-cells operate over a range of membrane potentials, we studied outward currents in islet cells from a range of holding potentials and found no differences between currents elicited from −90, −70, or −50 mV. Steady-state inactivation protocols (over 15 sec) showed sustained currents displaying a half-maximal voltage sensitivity ($V_{1/2}$) of −32.47 ± 1.53 mV (n = 12).

Consistent with its ability to inhibit Kv and K_{Ca} channels far more potently than K_{ATP} channels (12, 31), TEA (20 mM) inhibited outward K^+ currents from HIT-T15 and rat islet cells by 85.5 ± 2.7% (n = 9; $P < 0.001$) and 87.9 ± 1.2% (n = 11; $P < 0.001$), respectively (Fig. 5). The effect of TEA was reversible upon washing after

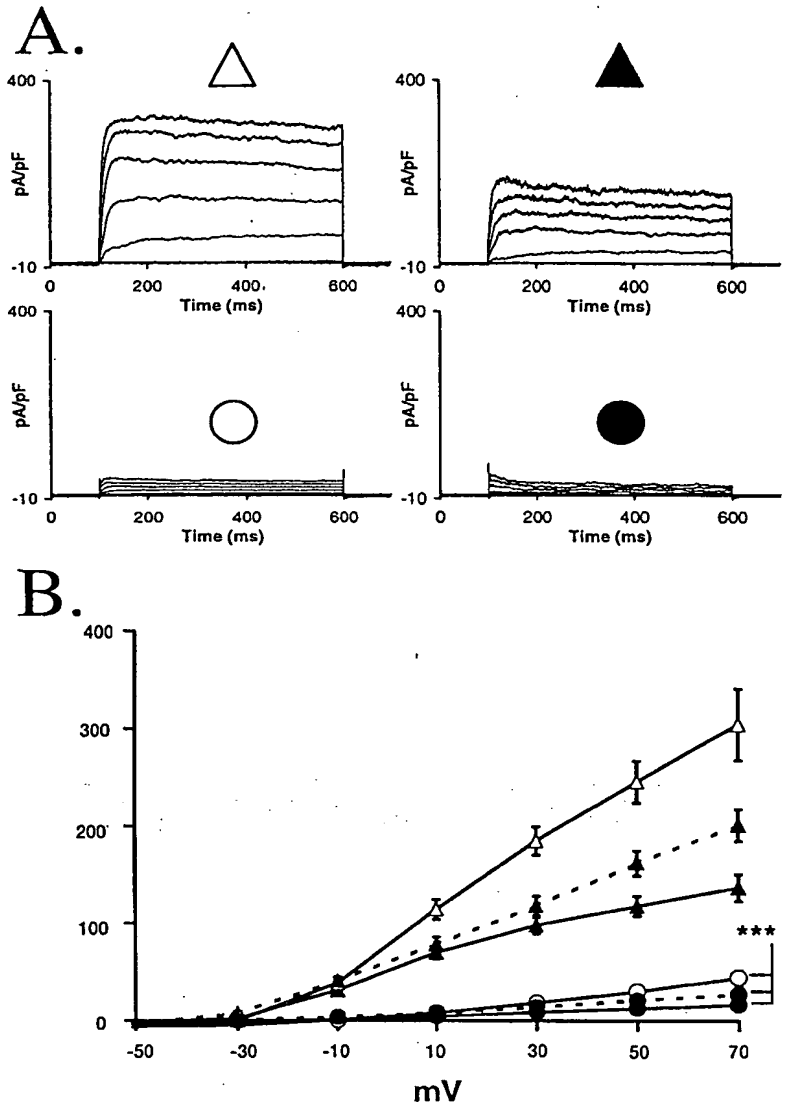


Fig. 5. β-Cell I_{DR} Is Blocked by TEA

Outward K^+ currents were recorded by depolarizing with a series of 500-msec pulses from a holding potential of −70 mV in 20-mV increments to a maximal depolarization to 70 mV. Data were normalized to cell capacitance. Representative traces from a typical HIT-T15 cell (*open marks*) and rat islet cell (*black marks*) are shown under control conditions (*triangles*) and in the presence of 20 mM TEA (*circles*) in panel A. In panel B, the current-voltage relationship of maximum sustained current was plotted for both HIT-T15 cells (*open marks*) and rat islet cells (*black marks*). At more physiological temperatures (31–33 C, *dashed line*), sustained outward currents were moderately larger and also were largely blocked by 20 mM TEA. ***, $P < 0.001$ compared with controls.

exposures of as long as 2 h, similar to the exposures used for the above insulin secretion studies (data not shown). The biophysical and pharmacological properties of these currents most closely resembled those mediated by members of the Kv1, 2, and 3 families, but not the Kv4 family (37–39). Outward currents from rat islet cells at more physiological temperatures (31–33 C) were somewhat larger at the end of a 500-msec depolarization to 70 mV from −70 mV. However, current inhibition by 20 mM TEA (86.4 ± 1.2%, n = 9: $P < 0.001$) was not significantly different compared with room temperature.

Effect of Kv and K_{Ca} Channel Antagonists on Insulin Secretion

To investigate whether specific Kv or K_{Ca} channels contribute to the regulation of insulin secretion, exper-

iments were performed using selective channel antagonists. Margatoxin (100 nM), which inhibits Kv1.3 and 1.6 with an IC_{50} of 30 pM and 5 nM, respectively (40), did not effect insulin secretion from either HIT-T15 cells or rat islets (Table 2). Dendrotoxin (200 nM), an inhibitor of both Kv1.1 and 1.2 channels with an IC_{50} of 20 nM (41, 42), did enhance GSIS from HIT-T15 cells (Table 2) accompanied by a 26.3 ± 9.7% (n = 7; $P <$ 0.001) reduction in I_{DR}, but did not enhance insulin secretion from rat islets. This is consistent with our ability to detect mRNA transcripts for Kv1.1 and variable but low Kv1.1 protein in HIT-T15 cells, but not rat islets. Specific antagonists are not available against cloned Kv1.4 channels, the other Kv1 family member that was detected. However, heterotetrameric channels formed from this subunit are insensitive to TEA (41) and are therefore less likely contributors to TEA's insulinotropic effect. Because no specific antagonists to Kv2 family channels are commercially available, this characterization was limited to antagonists of Kv1 channel family members.

Because both large- and small-conductance K_{Ca} currents have been detected in insulin-secreting cells (43–48), we investigated the effect of K_{Ca} channel antagonists on GSIS from rat islets. Neither the small conductance K_{Ca} antagonist apamin (200 nM) nor the large- and medium-conductance K_{Ca} antagonist iberiotoxin (100 nM) had a significant effect on GSIS from rat islets compared with controls (Table 2). However, this does not rule out the possibility that an apamin-insensitive small-conductance K_{Ca} current may have a role in regulating insulin secretion (49, 50).

Effect of Dominant-Negative Knockout of Kv1 and 2 Channels on β-Cell I_{DR}

To further investigate the role of the Kv1 and 2 family channels in mediating β-cell I_{DR}, we used a recombinant adenovirus approach to express dominant-negative Kv1 (AdKv1.4N) and 2 (AdKv2.1N) channel subunits. Mutation or truncation involving all or part of the pore-forming loop results in nonfunctional subunits that can coassemble with and eliminate ion flow through endogenous channels of the same family. Similar approaches have been used to study and identify subunit assembly of native Kv channels (24, 51, 52).

Expression of the Kv1.4N subunit in HIT-T15 cells and rat islet cells decreased I_{DR} by 26.8 ± 5.9% (n = 14; $P < 0.05$) and 22.3 ± 5.3% (n = 8; $P < 0.05$), respectively, compared with controls (Fig. 6). Expression of Kv2.1N reduced I_{DR} in HIT-T15 cells and rat islets cells to a far greater extent (72.9 ± 2.9%; n = 24; $P < 0.001$ and 61.6 ± 3.2%; n = 22; $P < 0.001$, respectively) compared with enhanced green fluorescent protein (EGFP)-expressing controls (Fig. 7). TEA (20 mM) further reduced outward K^+ currents in cells expressing Kv2.1N, eliminating a total of 94.3 ± 1.8% (n = 7; $P < 0.001$) (HIT-T15) and 86.9 ± 1.8% (n = 11; $P < 0.001$) (rat islet cells) of I_{DR} compared with EGFP controls (Fig. 7). Remaining currents in Kv2.1N-expressing rat islet cells after the addition of 20 mM TEA resembled A currents mediated by cloned Kv1.4 and could be inactivated by holding at −50 mV, a protocol known to inactivate A currents (53) (Fig. 8). These results suggest that the Kv1 and Kv2 channel families contribute approximately 20–30% and about 60–70% of the I_{DR} in insulin-secreting cells, respectively, potentially accounting for 80–100% of total I_{DR} observed under the present conditions. Steady-state inactivation of K^+ currents recorded from rat islet cells was unchanged by the expression of the Kv1.4N or Kv2.1N constructs, showing no differences in voltage sensitivity of the inactivating portion of the remaining currents with $V_{1/2}$ values of −33.6 ± 1.6 and −37.7 ± 1.7 mV (n = 4 and 9).

Effect of Dominant-Negative Knockout of Kv1 and Kv2 Family Channels on Insulin Secretion

Isolated islets were infected *in vitro* with AdKv1.4N, AdKv2.1N, or AdEGFP (control). Coexpression of EGFP allowed visualization of infected cells and estimation of infection efficiency. Laser confocal microscopy (not shown) and our previous studies (54) have shown that infection efficiencies of 30–50% are typical and cells within the islet core can be infected. Expression of Kv1.4N in rat islets had no effect on basal insulin secretion but significantly enhanced GSIS compared with control (0.031 ± 0.004 to 0.043 ± 0.007 ng/islet/h, n = 12; $P < 0.05$) (Fig. 9A). Likewise, expression of Kv2.1N in rat islets did not effect basal insulin secretion and caused a much larger enhancement of GSIS compared with control (0.044 ± 0.009 to

Table 2. Effect of Kv1 and K_{Ca} Specific Antagonists on Glucose-Stimulated Insulin Secretion from HIT-T15 Cells and Rat Islets

Antagonist	Channels Blocked	HIT-T15 Cells (ng/ml/2h)		Rat Islets (ng/islet/h)	
		Control	Drug	Control	Drug
α-Dendrotoxin (200 nM)	Kv1.1, 1.2	0.39 ± 0.03	0.51 ± 0.03[a]	0.15 ± 0.04	0.15 ± 0.02
Margatoxin (100 nM)	Kv1.3, 1.6	0.57 ± 0.07	0.49 ± 0.04	0.09 ± 0.01	0.10 ± 0.01
Apamin (200 nM)	SK_{Ca} channels			0.25 ± 0.03	0.27 ± 0.05
Iberiotoxin (100 nM)	BK_{Ca} channels			0.11 ± 0.02	0.10 ± 0.01

[a] $P < 0.05$ compared to control.

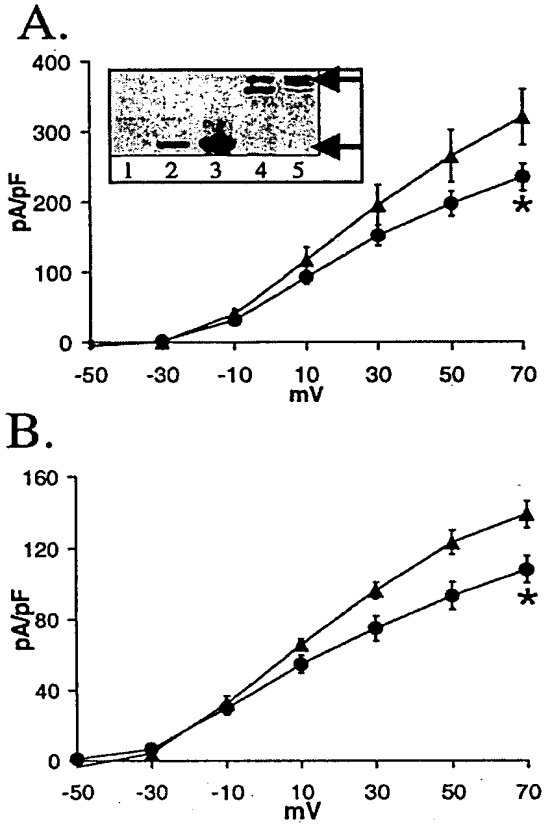


Fig. 6. Kv1.4N Expression Reduces β-Cell I_{DR}

Current-voltage relationships were obtained from HIT-T15 cells (A) and rat islet cells (B) expressing control EGFP (*triangles*) or the dominant-negative Kv1.4N construct (*circles*). *Inset*, Western blotting for the Kv1.4N construct showed expression of the truncated protein in Kv1.4N-GW1H-transfected (2) and AdKv1.4N-infected (3) HIT-T15 cells; only the full-length protein was detected in Kv1.4-GW1H-transfected (4) or AdKv1.4-infected (5) cells. Upon longer exposure, endogenous Kv1.4 would be detectable in control lysates (1). *, $P < 0.05$ compared with controls.

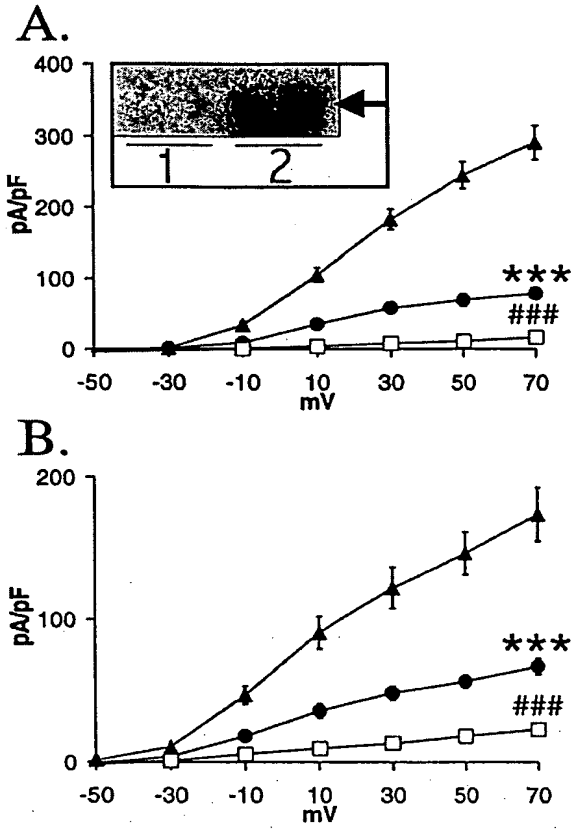


Fig. 7. Kv2.1N Expression Reduces β-Cell I_{DR}

Current-voltage relationships were obtained from HIT-T15 cells (A) and rat islet cells (B) expressing control EGFP (*triangles*) or the dominant-negative Kv2.1N construct (*circles*). Outward currents in cells expressing Kv2.1N could still be reduced by addition of 20 mM TEA (*open squares*). *Inset*, Northern blotting for the Kv2.1N transcript showed expression in AdKv2.1N-infected (2) HIT-T15 cells (n = 2); no transcript was detected in control-infected (1) cells (n = 2). ***, $P < 0.001$ compared with controls; and ###, $P < 0.001$ compared with Kv2.1N-expressing cells.

0.070 ± 0.018 ng/islet/h, n = 9; $P < 0.001$) (Fig. 9B). These results appear to be in good agreement with our electrophysiological observations, providing further evidence for a link between enhanced insulin secretion and reduction of I_{DR}.

DISCUSSION

Repolarization of pancreatic β-cells after a glucose-induced depolarization is mediated by a voltage-dependent outward K^+ current, which assists in closure of voltage-dependent Ca^{2+} channels, thereby modulating insulin secretion (5, 11–14). Accordingly, the general I_{DR} inhibitor TEA enhances glucose-stimulated $[Ca]_i$ oscillations and insulin secretion (11–13, 16, 31). Consistent with an important role for these currents in β-cells, we found that 20 mM TEA reduced I_{DR} (by 85–90% at both room temperature and near-physiological temperature) and enhanced glucose-stimulated insulin secretion (~2- to 4-fold) in both HIT-T15 cells and isolated rat islets. As expected, since β-cell I_{DR} currents are postulated to activate only after glucose induced depolarization, TEA had no insulinotropic effect in the absence of stimulatory glucose. The ability of TEA to block I_{DR} and enhance glucose-dependent insulin secretion suggests that repolarizing K^+ channels underlie I_{DR}. However, the effects of TEA do not resolve which K^+ channels are responsible for I_{DR} in β-cells.

For a number of reasons, it is unlikely that TEA exerts its glucose-dependent insulinotropic effect by inhibiting K_{ATP} channels. Unlike K_{ATP} antagonists such

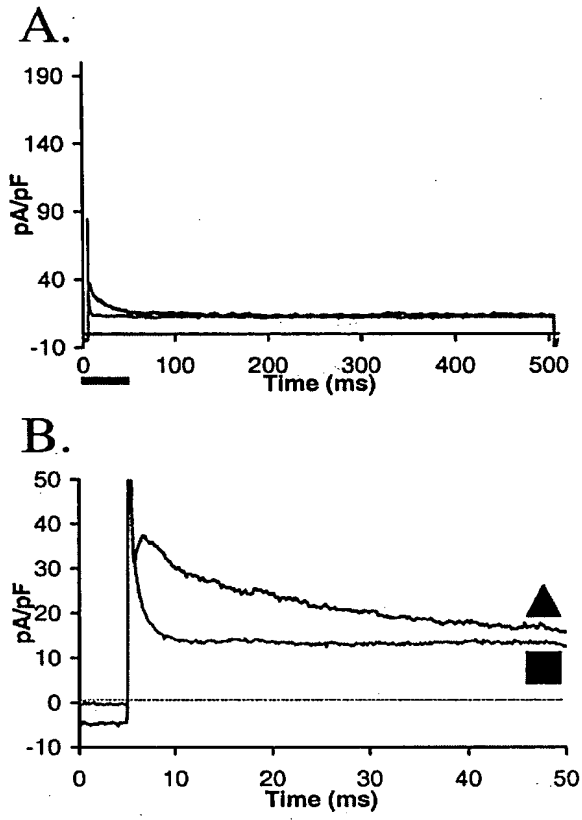


Fig. 8. Outward K^+ Currents in Kv2.1N-Expressing Cells Exposed to TEA

Remaining outward currents in AdKv2.1N-infected rat islet cells exposed to TEA (20 mM) were small and displayed an A current component when depolarized to 30 mV from a holding potential of −90 mV (*triangle*). Holding the cells at a more positive potential (−50 mV; *square*) before depolarization did not affect sustained currents (A), but dramatically reduced the Kv1.4-like A current component (B). Each trace is an average of recordings from eight AdKv2.1N-infected rat islet cells; the time represented by the *black bar* in panel A is shown on an expanded scale in panel B. The very fast component (within 5 msec of depolarization) results from uncompensated capacitance transient, and the small differences in initial holding current result from the different holding potentials.

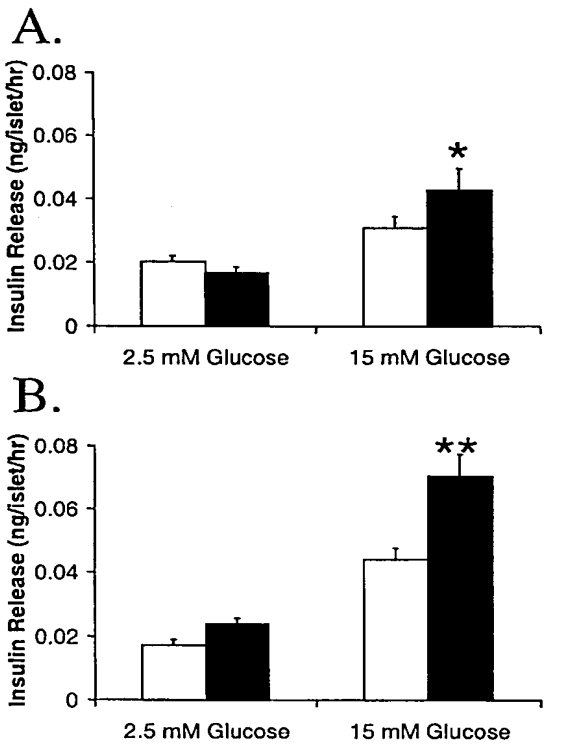


Fig. 9. Kv1.4N and Kv2.1N Expression Enhances GSIS from Rat Islets

Insulin secretion from AdKv1.4N (panel A, *black bars*) and AdKv2.1N (B, *black bars*)-infected rat islets was enhanced compared with controls (*white bars*). These dominant-negative subunits enhanced insulin secretion only in the presence of stimulatory glucose, while no effect was observed under nonstimulatory conditions. *, $P < 0.05$; and **, $P < 0.01$ compared with controls.

as glyburide, TEA (20 mM) did not enhance unstimulated insulin secretion (Figs. 2 and 3) (9, 10). In fact, the combination of TEA and glyburide enhanced insulin secretion to a greater degree than either alone, suggesting separate targets. Moreover, the glucose-dependent insulinotropic effect of TEA was observable at concentrations far lower than the published EC_{50} for K_{ATP} channels (Fig. 1C). Finally, in the presence of high glucose, the majority of K_{ATP} channels are closed, owing to an increase in the ATP:ADP ratio (11, 55).

Glyburide enhances insulin secretion from rodent islets with an EC_{50} of 0.5 nM (56), while human islets bind glyburide with a dissociation constant (K_d) of 1 nM (34). Here, a glyburide concentration of 10 nM stimulated a 2-fold increase in insulin secretion from isolated rat islets in the absence of stimulatory glucose. TEA enhanced glyburide-stimulated insulin release, indicating that membrane depolarization is sufficient to allow TEA's insulinotropic effect. The inability of TEA to significantly enhance rat islet insulin secretion stimulated by 2 μM glyburide (Fig. 3A) may result from nonspecific effects of this high dose of glyburide on other cell types within the islet, a problem that would not be present in a homogenous insulinoma cell line. Interestingly, in the presence of stimulatory glucose, the effects of glyburide or the phosphodiesterase inhibitor IBMX were enhanced by TEA (Fig. 3, C and D), suggesting that TEA-like drugs may be used in combination with K_{ATP} or PKA pathway agonists for a greater insulinotropic effect.

It is conceivable that Ca^{2+}-sensitive K^+ currents mediate the effects of TEA in our studies. Indeed K_{Ca} currents have been detected in insulin-secreting cells; however, reports regarding the pharmacological identification of these currents and their contri-

bution to glucose-induced electrical activity are conflicting (12, 30, 44–46, 48–50, 57–59). There is little functional evidence supporting a major role for K_{Ca} channels in regulating insulin secretion, and we were unable to detect K_{Ca} protein or an insulinotropic effect of general K_{Ca} channel antagonists (100 nM iberiotoxin and 200 nM apamin) in rat islets (Table 2). It is possible, nevertheless, that an apamin-insensitive small-conductance K_{Ca} current, possibly mediated by SK1 (60), can modulate insulin secretion (45, 49, 50).

Although it seems clear that Kv channels are mediators of β-cell membrane repolarization, a role for specific channels in mediating I_{DR} has not been established. Since Kv channels consist of homo- or heterotetrameric proteins from the same family (17, 23, 25, 29), we chose to express truncated subunits lacking the pore-forming region to selectively knock out functional channels in a family-specific manner. Similar approaches have been used to study and identify α-subunit assembly of native Kv channels (24, 51, 52). In our study, the dominant-negative Kv1.4N and Kv2.1N constructs inhibited outward K^+ currents when coexpressed with wild-type channels of the same family in HIT-T15 cells, but did not inhibit currents resulting from different channel families (members of the Kv1, 2, 3, and 4 channel families were tested; data not shown).

Expression of Kv2.1N in HIT-T15 cells or rat islet cells had a dramatic effect on I_{DR}, reducing it by approximately 70 and 60%, respectively. This correlated with an approximately 60% increase in GSIS from Kv2.1N infected islets compared with EGFP-expressing controls. Supported by the fact that the EC_{50} for the insulinotropic effect of TEA is within the range reported for Kv2.1's IC_{50} for block by TEA (61–63), our data suggest an important role for the Kv2 family in insulin secretion. Kv2.1 protein was detected at levels comparable to the rat brain control in both the insulinoma cell lines and rat islets. This is consistent with previous studies showing high-level protein expression of Kv2.1 in βTC3-neo insulinoma cells and Kv2.1 mRNA in insulin-secreting cells (5, 11). Transcripts for Kv2.2, the only other Kv2 family member that forms functional channel pores, were not detected. Kv2.1N expression did not enhance insulin secretion to the same degree as seen with TEA and may be explained in a number of ways. The insulinotropic effect of TEA was measured in response to an acute application of the drug, whereas the effect of Kv2.1N expression was measured after a more chronic expression protocol (2 days) that may have led to changes in the machinery controlling insulin secretion. In addition, our adenoviral expression of the Kv2.1N construct was limited to approximately 50% of the cells. Infection of rat islets with control EGFP virus decreased basal insulin secretion and reduced insulin secretion induced by glucose. Although the degree of insulin secretion enhancement by Kv2.1N expression was compared with EGFP controls, it is conceivable that Kv2.1N might contribute additional effects on insulin secretion independent of I_{DR} reduction. To minimize the possible effects of differential expression efficiency between control and experimental groups, islets were infected with equal numbers of viral particles and inspected for qualitatively similar levels of EGFP expression. Finally, it is still uncertain whether the relationship between I_{DR} reduction and enhancement of GSIS is linear, meaning that a reduction in I_{DR} greater than 60–70% may be required for a 2- to 4-fold increase in insulin secretion to occur.

Expression of Kv1.4N in HIT-T15 or rat islet cells reduced I_{DR} by approximately 30 and 20%, respectively, and increased GSIS from rat islets by about 40% compared with EGFP controls. Of the Kv1 channel family, Kv1.6 protein was detected at high levels in rat islets, while Kv1.4 protein was detected at high levels in rat islets and the insulinoma cell lines HIT-T15 and βTC-6f7. Kv1.2 protein was detected at low levels in rat islets, and Kv1.1 protein was detected variably at low levels in HIT-T15 cells. We did not examine the protein expression of Kv1.5 or 1.7, as neither was detectable in insulin-secreting cells by RT-PCR, and both are known to be insensitive to TEA. Variable detection of Kv1.1 in HIT-T15 cells is consistent with the ability of Dendrotoxin to reduce I_{DR} and enhance insulin secretion in these cells. Our results suggest a minimal contribution of homotetrameric Kv1.6 or Kv1.4 channels to the insulinotropic effect of TEA since the former is sensitive to Margatoxin and the latter is insensitive to TEA. However, heterotetrameric channels containing these subunits cannot be ruled out since heterotetrameric channels do not necessarily possess the pharmacological sensitivities of their constituent subunits (29). Also, the presence of regulatory β-subunits, channel phosphorylation, and the channels oxidative state are known to significantly alter channel pharmacology and kinetics (27, 28, 64–67). We did observe a small A current component in Kv2.1N-expressing rat islet cells in the presence of 20 mM TEA that was inactivated by holding the cell at −50 mV. This provides confirmatory evidence for the presence of Kv1.4-containing channels but suggests a limited role for them under normal conditions.

Current type 2 diabetes treatments aimed at enhancing insulin secretion are limited to the sulfonylurea drugs, which act in a glucose-independent manner. This is because their mechanism involves inhibition of $K_{ir}6.2$ through an interaction with the associated SUR1, depolarizing the cell, and triggering influx of Ca^{2+} and ultimately insulin secretion. Because TEA acts in a glucose-dependent fashion, enhancing β-cell depolarization rather than initiating it, drugs acting at TEA's specific target may be considered useful therapies that could also be expected to enhance the insulinotropic effect of K_{ATP} or PKA pathway agonists. In this study we identified high-level expression of Kv1.4, 1.6, and 2.1 in rat islets and have used an adenoviral approach to functionally knock

out these channels in isolated islets. Dominant-negative knockout of Kv2.1 enhanced insulin secretion by 60% in a glucose-dependent manner, while knockout of the Kv1 channel family members had a similar, but lesser, effect. It seems clear, however, that Kv2.1, and potentially members of the Kv1 channel family, may represent novel targets for the treatment of type 2 diabetes.

MATERIALS AND METHODS

Cell Culture and Islet Isolation

HIT-T15 cells, a gift from R. P. Robertson (Pacific NW Research Institute, Seattle, WA), passage 80–95, were cultured in Roswell Park Memorial Institute (RPMI) 1640 medium supplemented with 10% FBS, 1% L-glutamine, and 1% penicillin-streptomycin. Islets of Langerhans were isolated from male Wistar rats, 250–350 g, by perfusion of the pancreas through the common bile duct with 10 ml of a collagenase solution (10 mg/100 g body wt) and incubation of the excised pancreas with shaking at 37 C. The digestion was washed, filtered through 355 μm mesh, and separated on a density gradient created by resuspending the pellet in histopaque-1077 (Sigma, St. Louis, MO) and layering on serum-free media [low-glucose (LG)-RPMI 1640 described below without serum). Islets were collected from the interphase and further purified from contaminating single cell types by sedimentation. Isolated islets were cultured in LG-RPMI 1640 (7.5% FBS, 1% penicillin/streptomycin, 0.25% HEPES, and 2.5 mM glucose) at 37 C and 5% CO_2.

Insulin Secretion Studies

Twenty islets per well were plated in 24-well plates with LG-RPMI 1640 for insulin secretion studies. Twenty-four to 48 h after isolation, islets were washed and LG-RPMI 1640 was replaced by 2 ml of experimental media. Experimental media consisted of either LG-RPMI 1640 or high glucose (HG)-RPMI 1640 (15 mM glucose) with or without various experimental agents (see figures).

For HIT-T15 cell studies, cells were plated in 12-well plates at 5×10^5 cells per well. Forty-eight hours after plating, HIT-T15 cells were washed with, and preincubated for 2x 30 min in, Krebs Ringer bicarbonate (KRB) buffer (115 mM NaCl, 5 mM KCl, 24 mM $NaHCO_3$, 2.5 mM $CaCl_2$, 1 mM $MgCl_2$, 10 mM HEPES, and 0.1% BSA). After preincubation, cells were washed with KRB buffer and then incubated in 1 ml of KRB buffer alone or with 10 mM glucose with and without experimental agents (see figures).

All secretion studies were performed for 2 h at 37 C and 5% CO_2, after which media samples were taken and centrifuged at 700 $\times$ *g*. RIAs were performed using a Rat Insulin RIA Kit (Linco Research, Inc., St. Charles, MO). Each experiment was performed with an n value of at least 8 in at least three separate experiments, and data were normalized to an unstimulated control to account for variation between preparations and are expressed as nanograms/islet/h or nanograms/ml/2 h. Data were analyzed with Student's *t* test or Wilcoxon matched pairs test as appropriate. Dose-response curves and EC_{50} values for insulin secretion studies were generated using PRISM software (GraphPad Software, Inc., San Diego, CA).

Dominant-Negative Kv Channel Constructs and Adenoviral Vectors

E1-deleted recombinant adenovirus shuttle vectors expressing a C-terminal truncated Kv1.4 subunit (AdKv1.4N) or enhanced green fluorescent protein (AdEGFP-RSV) alone under the control of the rous sarcoma virus promoter was provided by Dr. Roger J. Hajjar (Cardiovascular Research Center and Heart Failure Transplantation Center, Massachusetts General Hospital, Harvard Medical School, Boston, MA). Recombinant adenoviruses expressing a C-terminal truncated Kv2.1 subunit (AdKv2.1N) or EGFP alone (AdEGFP-CMV) under the control of the cytomegalovirus promoter were prepared by CRE-lox recombination (68). All of these adenovirus constructs coexpress EGFP with the gene of interest to facilitate the identification of infected cells. Adenoviruses were amplified by passage in HEK 293 cells or CRE-8 cells (for viruses constructed by CRE-lox recombination). Infected cells were resuspended and lysed in 10 mM Tris, 1 mM $MgCl_2$, pH 8.0 [1 mM freeze-thaw media (FT)] and purified by centrifuging the lysate on a gradient created by layering 3 ml each of 1.20 g/ml, 1.33 g/ml, and 1.45 g/ml CsCl in 1 mM FT at 27,000 rpm for 2 h in a SW41-T1 rotor (Beckman Coulter, Inc., Fullerton, CA). Resultant bands were removed and dialyzed overnight against 1 mM FT and 10% glycerol and stored at −70 C until use.

Infection of isolated rat islets was performed in 24-well plates with either 20 (insulin secretion studies) or 50 (electrophysiological studies) islets per well on the day of isolation. Infection of HIT-T15 cells for electrophysiological studies (AdKv2.1N only) was performed in 35-mm dishes seeded 24 h previously with 5×10^5 cells per dish. Islets or HIT-T15 cells were cultured in 0.5 ml of normal media with 1×10^{10} virus particles/ml for 2 h at 37 C and 5% CO_2 after which 1.5 ml of LG-RPMI 1640 were added. Forty-eight hours later, islets or HIT-T15 cells were examined under UV light to detect the expression of EGFP. Insulin secretion studies, electrophysiological studies, RNA isolation, or protein isolation was carried out 48 h post infection.

For HIT-T15 cell electrophysiological studies, a wild-type Kv1.4 or a Kv1.4N construct (in the GW1H plasmid; provided by Dr. Hajjar) was expressed by transfection with Lipofectamine (Life Technologies, Inc., Gaithersburg, MD) as per instructions of the manufacturer. This plasmid was cotransfected with the pEGFP plasmid (CLONTECH Laboratories, Inc. Palo Alto, CA) that expresses EGFP as a marker for transfection. Control cells were transfected with pEGFP alone.

Electrophysiological Studies

Islets were washed in and incubated with PBS and 0.2 mM EDTA with 1.5% trypsin for 11 min, followed by mechanical dispersion and plating of single-islet cells overnight in LG-RPMI 1640 in 35-mm culture dishes. Cells were voltage clamped in the whole-cell configuration using an EPC-9 amplifier and Pulse software (Heka Electronik, Lambrecht, Germany). Electrical identification of β-cells using a current clamp was not possible due to the intracellular solution required to measure I_{DR} currents; however, the majority of islet cells (~70% or more) are β-cells, and all electrophysiological experiments were confirmed in a clonal β-cell line (HIT-T15). HIT-T15 cells were trypsinized and replated in 35-mm dishes 24 h before electrophysiological studies. Patch pipettes were prepared from 1.5-mm thin-walled borosilicate glass tubes using a two-stage micropipette puller (Narishige, Tokyo, Japan). Pipettes were heat polished and typically had a tip resistance of 3–6 MΩ when filled with intracellular solution containing (in mM): KCl, 140; $MgCl_2 \cdot 6\ H_2O$, 1; EGTA, 1; HEPES, 10; MgATP 5 (pH 7.25) with KOH. The bath solution contained (in mM): NaCl, 140; $CaCl_2$, 2; KCl, 4; $MgCl_2 \cdot 6\ H_2O$, 1; HEPES, 10 (pH 7.3) with NaOH. All electrophysiological measurements reported were made at room temperature (22–24 C) and normalized to cell capacitance unless stated otherwise. For experiments at 31–33 C, temperature was maintained with an Olympus America Inc. temperature control unit (Melville, NY) and continuous perfusion with warmed solutions. Outward currents were elicited with a 500-msec

depolarization in steps of 20 mV to +70 mV from a holding potential of −70 mV. Outward currents were also compared from holding potentials of −90, −70, and −50 mV using 500-msec depolarizing pulses to 30 mV. To minimize variation, maximum sustained current was determined from a third degree polynomial function fit to the final 25 msec of the 500-msec depolarizing pulse.

The voltage dependence of steady state inactivation was investigated by holding the cells at potentials from −80 to 30 mV for 15 sec followed by a 5-msec prepulse to −70 and a 500-msec depolarization to 30 mV to elicit outward currents. Steady state inactivation curves were fit with a Boltzman function: $I/I_{max} = 1/[1 + \exp([V - V_{1/2}]/s)]$ where $V_{1/2}$ is the voltage at which half the channels are inactivated, and s is the slope of the curve. For pharmacological studies, the drug was applied by perfusion for at least 5 min before recording. Outward currents at the end of the 500-msec depolarizing pulse were compared using the *t* test.

RNA Analysis

Total RNA was obtained from rat islets (24–48 h after isolation), rat brain, and HIT-T15 cells using Trizol (Life Technologies, Inc.) as per the manufacturer's instructions. RT-PCR was performed on 1 μg of total RNA using a GeneAmp RNA PCR kit (Perkin-Elmer Corp., Branchburg, NJ) according to the manufacturer's instructions. PCR primers used were designed to conserved sequences of rat Kv1.1 [Forward (F): 5′-AAGGATCCGTCATTGTGTCC-3′; Reverse (R): 5′-AAAGGCCTAAACATCGGTCAG-3′], Kv1.2 (F: 5′-GTAAAGCACACTTCTCAAGCCCC-3′; R: 5′-CCTCCCGAAACATCTCAATTGC-3′); Kv1.3 (F: 5′-GAGATCCGCTTTTACCAGCTGGG-3′; R: 5′-CATGATATTTCTGGAGAAGG-3′); Kv1.4 (F: 5′-GATAGCCATTGTGTCCGTCCTGG-3′; R: 5′-GGCACACAGGGACCCGACAATC-3′); Kv1.5 (F: 5′-CTGAGAGGGAGAGAGGCAGGG-3′; R: 5′-GCAGCTCCTGAGGCATAGGG-3′); Kv1.6 (F: 5′-GTTGGTGATCAACATCTCCGGG-3′; R: 5′-GGCCGCCTTGCTGGGACAGG-3′); Kv1.7 (mouse) (F: 5′-TCTCCGTACTCGTCATCCGG-3′; R: 5′-AAATGGGTGTCCACCCGGTC-3′); Kv2.1 (F: 5′-CGAGGAGCTGAAGCGGGAGG-3′; R: 5′-GGAAGATGGTGACGTAGTAGGG-3′); and Kv2.2 (F: 5′-GGATGCCTTTGCTAGAAGTATGG-3′; R: 5′-CGCTGGCACTGTCAGGTTGC-3′). PCR was also performed on water blank controls containing no cDNA template and rat brain cDNA as a positive control. PCR was performed with 35 cycles of 94 C for 30 sec, 60 C for 35 sec, and 72 C for 45 sec followed by a 10-min extension at 72 C. PCR products of the expected size were excised from an 1.2% low melt agarose gel and ligated into the pCR2.1 vector and sequenced using the universal M13 reverse primer. Resulting sequences were subjected to analysis by NCBI Blast (NCBI, Bethesda, MD) and nucleotide and amino acid identity analysis with MacDNASIS (Hitachi Software, San Francisco, CA).

Northern analysis was used to detect expression of mRNA transcripts for Kv2.1N in total RNA (7.5 μg) from AdKv2.1N- or AdEGFP-infected HIT cells as described previously (69). Probes were generated by random priming (Random Primers DNA Labeling System, Life Technologies, Inc.) of Kv2.1N cDNA and incorporation of P^{32}-dCTP. Blots were washed twice by shaking in room temperature 0.1% SDS/2×SSC followed by a 30-min wash in 0.1% SDS/0.1× SSC at 55 C. Blots were exposed overnight to X-OMAT AR film (Eastman Kodak Co., Rochester, NY).

Protein Analysis

Immunoblotting of Kv channel proteins was performed as previously described (70, 71). Briefly, the islets were washed in ice-cold PBS, solubilized in 2% SDS loading buffer, boiled for 10 min, and passed through a 23G needle. Fifty micrograms of the protein from each sample, determined by Lowry's method, were loaded and separated on a 10% polyacrylamide gel. The protein was transferred to PVDF-Plus (Fisher Scientific Ltd., Nepean, Ontario, Canada) membrane and immunodecorated with primary antibody or antibody-antigen solutions (diluted according to the supplier's instructions) for 1.5 h at room temperature. Primary antibodies were from Alomone Labs (Jerusalem, Israel) (Kv1.2, 1.3, 1.4, 1.6, 2.1) and Upstate Biotechnology, Inc. (Lake Placid, NY) (Kv1.1, 2.1). Primary antibodies were detected with appropriate secondary antibodies (sheep antimouse, 1:10,000; donkey antirabbit, 1:7,500; Amersham Pharmacia Biotech Ltd., Buckinghamshire, U.K.) for 1 h, and then visualized by chemiluminescence (ECL-Plus, Amersham Pharmacia Biotech Ltd.) and exposure of the filters to Kodak film (Eastman Kodak Co., Rochester, NY) for 5 sec to 10 min. At least three blots were performed for each protein investigated.

Acknowledgments

We thank Dr. Robert Hajjar (Harvard Medical School) for providing Kv1.4N plasmid and adenovirus vectors. Additionally, we thank Dr. Robert Tsushima (University of Toronto) for helpful discussion, the use of equipment, and critical reading of the manuscript; and Dr. Sabine Sewing (Eli Lilly) for helpful discussion.

Received January 31, 2001. Accepted May 8, 2001.

Address requests for reprints to: Michael B. Wheeler, Ph.D., or Peter H. Backx, D.V.M., Ph.D., University of Toronto, Department of Physiology, 1 Kings College Circle, Toronto, Ontario, Canada, M5S 1A8. E-mail: michael.wheeler@utoronto.ca or p.backx@utoronto.ca.

This research was supported by research grants to M.B.W. and P.H.B. from the Banting and Best Diabetes Centre (BBDC) and Eli Lilly & Co. (Indianapolis, IN). P.H.B. holds a Career Investigator Award from the Heart and Stroke Foundation of Ontario. P.E.M. was supported by studentships from the Department of Physiology, University of Toronto, and the BBDC/Novo Nordisk. S.R.S. was supported by an Institute of Medical Science Summer Studentship.

REFERENCES

1. Dagogo-Jack S, Santiago JV 1997 Pathophysiology of type 2 diabetes and modes of action of therapeutic interventions. Arch Intern Med 157:1802–1817
2. Straub SG, James RF, Dunne MJ, Sharp GW 1998 Glucose activates both K(ATP) channel-dependent and K(ATP) channel-independent signaling pathways in human islets. Diabetes 47:758–763
3. Aizawa T, Komatsu M, Asanuma N, Sato Y, Sharp GW 1998 Glucose action 'beyond ionic events' in the pancreatic β cell. Trends Pharmacol Sci 19:496–499
4. Ammala C, Kane C, Cosgrove KE, et al. 1997 Characterization of ion channels in stimulus-secretion coupling in pancreatic islets. Digestion 58 (Suppl 2):81–85
5. Dukes ID, Philipson LH 1996 K^+ channels: generating excitement in pancreatic β-cells. Diabetes 45:845–853
6. Newgard CB, McGarry JD 1995 Metabolic coupling factors in pancreatic β-cell signal transduction. Annu Rev Biochem 64:689–719
7. Rorsman P 1997 The pancreatic β-cell as a fuel sensor: an electrophysiologist's viewpoint. Diabetologia 40: 487–495
8. Tanabe K, Tucker SJ, Matsuo M, et al. 1999 Direct photoaffinity labeling of the Kir6.2 subunit of the ATP-sensitive K^+ channel by 8-azido-ATP. J Biol Chem 274: 3931–3933
9. Boyd III AE 1992 The role of ion channels in insulin secretion. J Cell Biochem 48:235–241

10. Dunne MJ, Cosgrove KE, Shepherd RM, Ammala C 1999 Potassium channels, sulphonylurea receptors and control of insulin release. Trends Endocrinol Metab 10: 146–152
11. Roe MW, Worley III JF, Mittal AA, et al. 1996 Expression and function of pancreatic β-cell delayed rectifier K^+ channels. Role in stimulus-secretion coupling. J Biol Chem 271:32241–32246
12. Fatherazi S, Cook DL 1991 Specificity of tetraethylammonium and quinine for three K channels in insulin-secreting cells. J Membr Biol 120:105–114
13. Henquin JC 1990 Role of voltage- and Ca2(+)-dependent K^+ channels in the control of glucose-induced electrical activity in pancreatic B-cells. Pflugers Arch 416: 568–572
14. Smith PA, Bokvist K, Arkhammar P, Berggren PO, Rorsman P 1990 Delayed rectifying and calcium-activated K^+ channels and their significance for action potential repolarization in mouse pancreatic β-cells. J Gen Physiol 95:1041–1059
15. Philipson LH, Rosenberg MP, Kuznetsov A, et al. 1994 Delayed rectifier K^+ channel overexpression in transgenic islets and β-cells associated with impaired glucose responsiveness. J Biol Chem 269:27787–27790
16. Eberhardson M, Tengholm A, Grapengiesser E 1996 The role of plasma membrane K^+ and Ca2+ permeabilities for glucose induction of slow Ca2+ oscillations in pancreatic β-cells. Biochim Biophys Acta 12:8367–8372
17. Christie MJ 1995 Molecular and functional diversity of K^+ channels. Clin Exp Pharmacol Physiol 22:944–951
18. Salinas M, de Weille J, Guillemare E, Lazdunski M, Hugnot JP 1997 Modes of regulation of shab K^+ channel activity by the Kv8.1 subunit. J Biol Chem 272: 8774–8780
19. Hugnot JP, Salinas M, Lesage F, et al. 1996 Kv8.1, a new neuronal potassium channel subunit with specific inhibitory properties towards Shab and Shaw channels. EMBO J 15:3322–3331
20. Stocker M, Kerschensteiner D 1998 Cloning and tissue distribution of two new potassium channel α-subunits from rat brain. Biochem Biophys Res Commun 248: 927–934
21. Salinas M, Duprat F, Heurteaux C, Hugnot JP, Lazdunski M 1997 New modulatory α subunits for mammalian Shab K^+ channels. J Biol Chem 272:24371–24379
22. Dilks D, Ling HP, Cockett M, Sokol P, Numann R 1999 Cloning and expression of the human kv4.3 potassium channel. J Neurophysiol 81:1974–1977
23. Xu J, Yu W, Jan YN, Jan LY, Li M 1995 Assembly of voltage-gated potassium channels. Conserved hydrophilic motifs determine subfamily-specific interactions between the α-subunits. J Biol Chem 270:24761–24768
24. Blaine JT, Ribera AB 1998 Heteromultimeric potassium channels formed by members of the Kv2 subfamily. J Neurosci 18:9585–9593
25. Koch RO, Wanner SG, Koschak A, et al. 1997 Complex subunit assembly of neuronal voltage-gated K^+ channels. Basis for high-affinity toxin interactions and pharmacology. J Biol Chem 272:27577–27581
26. Standen NB, Quayle JM 1998 K^+ channel modulation in arterial smooth muscle. Acta Physiol Scand 164:549–557
27. Rettig J, Heinemann SH, Wunder F, et al. 1994 Inactivation properties of voltage-gated K^+ channels altered by presence of β-subunit. Nature 369:289–294
28. Heinemann SH, Rettig J, Graack HR, Pongs O 1996 Functional characterization of Kv channel β-subunits from rat brain. J Physiol (Lond) 493:625–633
29. Po S, Roberds S, Snyders DJ, Tamkun MM, Bennett PB 1993 Heteromultimeric assembly of human potassium channels. Molecular basis of a transient outward current? Circ Res 72:1326–1336
30. Bokvist K, Rorsman P, Smith PA 1990 Block of ATP-regulated and Ca2(+)-activated K^+ channels in mouse pancreatic β-cells by external tetraethylammonium and quinine. J Physiol (Lond) 423:327–342
31. Bokvist K, Rorsman P, Smith PA 1990 Effects of external tetraethylammonium ions and quinine on delayed rectifying K^+ channels in mouse pancreatic β-cells. J Physiol (Lond) 423:311–325
32. Kulkarni RN, Wang ZL, Wang RM, Smith DM, Ghatei MZ, Bloom SR 2000 Glibenclamide but not other sulphonylureas stimulates release of neuropeptide Y from perifused rat islets and hamster insulinoma cells. J Endocrinol 165:509–518
33. Seri K, Sanai K, Kurachima K, Imamura Y, Akita H 2000 (R)-ACX is a novel sulfonylurea compound with potent, quick and short-lasting hypoglycemic activity. Eur J Pharmacol 389:253–256
34. Giannaccini G, Lupi R, Trincavelli ML, et al. 1998 Characterization of sulfonylurea receptors in isolated human pancreatic islets. J Cell Biochem 71:182–188
35. Gromada J, Holst JJ, Rorsman P 1998 Cellular regulation of islet hormone secretion by the incretin hormone glucagon-like peptide 1. Pflugers Arch 435:583–594
36. Kanno T, Suga S, Wu J, Kimura M, Wakui M 1998 Intracellular cAMP potentiates voltage-dependent activation of L-type Ca2+ channels in rat islet β-cells. Pflugers Arch 435:578–580
37. Perney TM, Kaczmarek LK 1991 The molecular biology of K^+ channels. Curr Opin Cell Biol 3:663–670
38. Philipson LH, Miller RJ 1992 A small K^+ channel looms large. Trends Pharmacol Sci 13:8–11
39. Rorsman P, Trube G 1986 Calcium and delayed potassium currents in mouse pancreatic β-cells under voltage-clamp conditions. J Physiol (Lond) 374:531–550
40. Garcia-Calvo M, Leonard RJ, Novick J, et al. 1993 Purification, characterization, and biosynthesis of margatoxin, a component of *Centruroides margaritatus* venom that selectively inhibits voltage-dependent potassium channels. J Biol Chem 268:18866–18874
41. Grissmer S, Nguyen AN, Aiyar J, et al. 1994 Pharmacological characterization of five cloned voltage-gated K^+ channels, types Kv1.1, 1.2, 1.3, 1.5, and 3.1, stably expressed in mammalian cell lines. Mol Pharmacol 45: 1227–1234
42. Hurst RS, Busch AE, Kavanaugh MP, Osborne PB, North RA, Adelman JP 1991 Identification of amino acid residues involved in dendrotoxin block of rat voltage-dependent potassium channels. Mol Pharmacol 40:572–576
43. Satin LS, Hopkins WF, Fatherazi S, Cook DL 1989 Expression of a rapid, low-voltage threshold K current in insulin-secreting cells is dependent on intracellular calcium buffering. J Membr Biol 112:213–222
44. Findlay I, Dunne MJ, Petersen OH 1985 High-conductance K^+ channel in pancreatic islet cells can be activated and inactivated by internal calcium. J Membr Biol 83:169–175
45. Kozak JA, Misler S, Logothetis DE 1998 Characterization of a Ca2+-activated K^+ current in insulin-secreting murine βTC-3 cells. J Physiol (Lond) 509:355–370
46. Findlay I, Dunne MJ, Ullrich S, Wollheim CB, Petersen OH 1985 Quinine inhibits Ca2+-independent K^+ channels whereas tetraethylammonium inhibits Ca2+-activated K^+ channels in insulin- secreting cells. FEBS Lett 185:4–8
47. Smith PA, Bokvist K, Arkhammar P, Berggren PO, Rorsman P 1990 Delayed rectifying and calcium-activated K^+ channels and their significance for action potential repolarization in mouse pancreatic β-cells. J Gen Physiol 95:1041–1059
48. Tabcharani JA, Misler S 1989 Ca2+-activated K^+ channel in rat pancreatic islet B cells: permeation, gating and blockade by cations. Biochim Biophys Acta 982:62–72
49. Lebrun P, Atwater I, Claret M, Malaisse WJ, Herchuelz A 1983 Resistance to apamin of the Ca2+-activated K^+ permeability in pancreatic β-cells. FEBS Lett 161:41–44

50. Gopel SO, Kanno T, Barg S, Eliasson L, Galvanovskis J, Renstrom E, Rorsman P 1999 Activation of Ca(2+)-dependent K(+) channels contributes to rhythmic firing of action potentials in mouse pancreatic β cells. J Gen Physiol 114:759–770
51. Ribera AB, Pacioretty LM, Taylor RS 1996 Probing molecular identity of native single potassium channels by overexpression of dominant negative subunits. Neuropharmacology 35:1007–1016
52. Johns DC, Marban E, Nuss HB 1999 Virus-mediated modification of cellular excitability. Ann NY Acad Sci 868:418–422
53. Hille B 2001 Potassium channels and chloride channels. In: Hille B, ed. Ionic channels of excitable membranes. Sunderland, MA: Sinauer Associates Inc.; 99–116
54. Chan CB, MacDonald PE, Saleh MC, Johns DC, Marban E, Wheeler MB 1999 Overexpression of uncoupling protein 2 inhibits glucose-stimulated insulin secretion from rat islets. Diabetes 48:1482–1486
55. Dukes ID, McIntyre MS, Mertz RJ, et al. 1994 Dependence on NADH produced during glycolysis for β-cell glucose signaling. J Biol Chem 269:10979–10982
56. Panten U, Burgfeld J, Goerke F, et al. 1989 Control of insulin secretion by sulfonylureas, meglitinide and diazoxide in relation to their binding to the sulfonylurea receptor in pancreatic islets. Biochem Pharmacol 38: 1217–1229
57. Light DB, Van Eenenaam DP, Sorenson RL, Levitt DG 1987 Potassium-selective ion channels in a transformed insulin-secreting cell line. J Membr Biol 95:63–72
58. Atwater I, Rosario L, Rojas E 1983 Properties of the Ca-activated K^+ channel in pancreatic β-cells. Cell Calcium 4:451–461
59. Lebrun P, Malaisse WJ, Herchuelz A 1983 Activation, but not inhibition, by glucose of Ca2+-dependent K^+ permeability in the rat pancreatic B-cell. Biochim Biophys Acta 731:145–150
60. Bond CT, Maylie J, Adelman JP 1999 Small-conductance calcium-activated potassium channels. Ann NY Acad Sci 868:370–378
61. Frech GC, VanDongen AM, Schuster G, Brown AM, Joho RH 1989 A novel potassium channel with delayed rectifier properties isolated from rat brain by expression cloning. Nature 340:642–645
62. Taglialatela M, VanDongen AM, Drewe JA, Joho RH, Brown AM, Kirsch GE 1991 Patterns of internal and external tetraethylammonium block in four homologous K^+ channels. Mol Pharmacol 40:299–307
63. Ikeda SR, Soler F, Zuhlke RD, Joho RH, Lewis DL 1992 Heterologous expression of the human potassium channel Kv2.1 in clonal mammalian cells by direct cytoplasmic microinjection of cRNA. Pflugers Arch 422:201–203
64. Jonas JC, Gilon P, Henquin JC 1998 Temporal and quantitative correlations between insulin secretion and stably elevated or oscillatory cytoplasmic Ca2+ in mouse pancreatic β-cells. Diabetes 47:1266–1273
65. Ruppersberg JP, Stocker M, Pongs O, Heinemann SH, Frank R, Koenen M 1991 Regulation of fast inactivation of cloned mammalian IK(A) channels by cysteine oxidation: Nature 352:711–714
66. Stephens GJ, Owen DG, Robertson B 1996 Cysteine-modifying reagents alter the gating of the rat cloned potassium channel Kv1.4. Pflugers Arch 431:435–442
67. Walaas SI, Greengard P 1991 Protein phosphorylation and neuronal function. Pharmacol Rev 43:299–349
68. Hardy S, Kitamura M, Harris-Stansil T, Dai Y, Phipps ML 1997 Construction of adenovirus vectors through Cre-*lox* recombination. J Virol 71:1842–1849
69. Salapatek AM, MacDonald PE, Gaisano HY, Wheeler MB 1999 Mutations to the third cytoplasmic domain of the glucagon-like peptide 1 (GLP-1) receptor can functionally uncouple GLP-1-stimulated insulin secretion in HIT-T15 cells. Mol Endocrinol 13:1305–1317
70. Huang X, Wheeler MB, Kang YH, et al. 1998 Truncated SNAP-25 (1–197), like botulinum neurotoxin A, can inhibit insulin secretion from HIT-T15 insulinoma cells. Mol Endocrinol 12:1060–1070
71. Wheeler MB, Sheu L, Ghai M, et al. 1996 Characterization of SNARE protein expression in β cell lines and pancreatic islets. Endocrinology 137:1340–1348

EXHIBIT K

1: Biochem Biophys Res Commun. 2001 May 11;283(3):549-53.

ELSEVIER SCIENCE FULL-TEXT ARTICLE

Downregulation of K(+) channel genes expression in type I diabetic cardiomyopathy.

Qin D, Huang B, Deng L, El-Adawi H, Ganguly K, Sowers JR, El-Sherif N.

Department of Veterans Affairs, New York Harbor Healthcare System, Brooklyn Campus, Brooklyn, New York, 11209, USA.

Type I diabetic cardiomyopathy has consistently been shown to be associated with decrease of repolarising K(+) currents, but the mechanisms responsible for the decrease are not well defined. We investigated the streptozotocin (STZ) rat model of type I diabetes. We utilized RNase protection assay and Western blot analysis to investigate the message expression and protein density of key cardiac K(+) channel genes in the diabetic rat left ventricular (LV) myocytes. Our results show that message and protein density of Kv2.1, Kv4.2, and Kv4.3 are significantly decreased as early as 14 days following induction of type I diabetes in the rat. The results demonstrate, for the first time, that insulin-deficient type I diabetes is associated with early downregulation of the expression of key cardiac K(+) channel genes that could account for the depression of cardiac K(+) currents, I(to-f) and I(to-s). These represent the main electrophysiological abnormality in diabetic cardiomyopathy and is known to enhance the arrhythmogenecity of the diabetic heart. The findings also extend the extensive list of gene expression regulation by insulin. Copyright 2001 Academic Press.

PMID: 11341759 [PubMed - indexed for MEDLINE]

characterize the protein. A starting material that can only be used to produce a final product does not have a substantial asserted utility in those instances where the final product is not supported by a specific and substantial utility. In this case none of the proteins that are to be produced as final products resulting from processes involving the claimed cDNA have asserted or identified specific and substantial utilities. The research contemplated by Applicants to characterize potential protein products, especially their biological activities, does not constitute a specific and substantial utility. Identifying and studying the properties of the protein itself or the mechanisms in which the protein is involved does not define a "real world" context of use. Note, because the claimed invention is not supported by a specific and substantial asserted utility for the reasons set forth above, credibility has not been assessed. Neither the specification as filed nor any art of record discloses or suggests any property or activity for the cDNA compounds such that another non-asserted utility would be well established for the compounds.

Claim 1 is also rejected under 35 U.S.C. § 112, first paragraph. Specifically, since the claimed invention is not supported by either a specific and substantial asserted utility or a well established utility for the reasons set forth above, one skilled in the art would not know how to use the claimed invention.

Example 10: <u>DNA Fragment encoding a Full Open Reading Frame (ORF)</u>

Specification: The specification discloses that a cDNA library was prepared from human kidney epithelial cells and 5000 members of this library were

sequenced and open reading frames were identified. The specification discloses a Table that indicates that one member of the library having SEQ ID NO: 2 has a high level of homology to a DNA ligase. The specification teaches that this complete ORF (SEQ ID NO: 2) encodes SEQ ID NO: 3. An alignment of SEQ ID NO: 3 with known amino acid sequences of DNA ligases indicates that there is a high level of sequence conservation between the various known ligases. The overall level of sequence similarity between SEQ ID NO: 3 and the consensus sequence of the known DNA ligases that are presented in the specification reveals a similarity score of 95%. A search of the prior art confirms that SEQ ID NO: 2 has high homology to DNA Ligase encoding nucleic acids and that the next highest level of homology is to alpha-actin. However, the latter homology is only 50%. Based on the sequence homologies, the specification asserts that SEQ ID NO: 2 encodes a DNA ligase.

Claim 1: An isolated and purified nucleic acid comprising SEQ ID NO: 2.

Analysis: The following analysis includes the questions that need to be asked according to the guidelines and the answers to those questions based on the above facts:

1) Based on the record, is there a "well established utility" for the claimed invention? Based upon applicant's disclosure and the results of the PTO search, there is no reason to doubt the assertion that SEQ ID NO: 2 encodes a DNA ligase. Further, DNA ligases have a well-established use in the molecular biology art based on this class of protein's ability to ligate DNA. Consequently the answer to the question is yes.

Note that if there is a well-established utility already associated with the claimed invention, the utility need not be asserted in the specification as filed. In order to determine whether the claimed invention has a well-established utility the examiner must determine that the invention has a specific, substantial and credible utility that would have been readily apparent to one of skill in the art. In this case SEQ ID NO: 2 was shown to encode a DNA ligase that the artisan would have recognized as having a specific, substantial and credible utility based on its enzymatic activity.

Thus, the conclusion reached from this analysis is that a 35 U.S.C. § 101 rejection and a 35 U.S.C. § 112, first paragraph, utility rejection should not be made.

Example 11: <u>Animals with Uncharacterized Human Genes</u>

Specification: Kidney cells from a patient with Polycystic Kidney (PCK) Disease have been used to make a cDNA library. From this library 8000 nucleotide "fragments" have been sequenced but not yet used to express proteins in a transformed host cell nor have they been characterized in any other way. The 50 longest fragments, SEQ ID NO: 1-50, respectively, have been used to make transgenic mice. None of the 50 lines of mice have developed Polycystic Kidney Disease to date. The asserted utility is the use of the mice to research human genes from diseased human kidneys. The disease is inheritable, but chromosomal loci have not yet been identified. Neither the absence or presence of a specific protein has been identified with the disease condition.

EXHIBIT S

The Sequence of the Human Genome

J. Craig Venter,[1]* Mark D. Adams,[1] Eugene W. Myers,[1] Peter W. Li,[1] Richard J. Mural,[1] Granger G. Sutton,[1] Hamilton O. Smith,[1] Mark Yandell,[1] Cheryl A. Evans,[1] Robert A. Holt,[1] Jeannine D. Gocayne,[1] Peter Amanatides,[1] Richard M. Ballew,[1] Daniel H. Huson,[1] Jennifer Russo Wortman,[1] Qing Zhang,[1] Chinnappa D. Kodira,[1] Xiangqun H. Zheng,[1] Lin Chen,[1] Marian Skupski,[1] Gangadharan Subramanian,[1] Paul D. Thomas,[1] Jinghui Zhang,[1] George L. Gabor Miklos,[2] Catherine Nelson,[3] Samuel Broder,[1] Andrew G. Clark,[4] Joe Nadeau,[5] Victor A. McKusick,[6] Norton Zinder,[7] Arnold J. Levine,[7] Richard J. Roberts,[8] Mel Simon,[9] Carolyn Slayman,[10] Michael Hunkapiller,[11] Randall Bolanos,[1] Arthur Delcher,[1] Ian Dew,[1] Daniel Fasulo,[1] Michael Flanigan,[1] Liliana Florea,[1] Aaron Halpern,[1] Sridhar Hannenhalli,[1] Saul Kravitz,[1] Samuel Levy,[1] Clark Mobarry,[1] Knut Reinert,[1] Karin Remington,[1] Jane Abu-Threideh,[1] Ellen Beasley,[1] Kendra Biddick,[1] Vivien Bonazzi,[1] Rhonda Brandon,[1] Michele Cargill,[1] Ishwar Chandramouliswaran,[1] Rosane Charlab,[1] Kabir Chaturvedi,[1] Zuoming Deng,[1] Valentina Di Francesco,[1] Patrick Dunn,[1] Karen Eilbeck,[1] Carlos Evangelista,[1] Andrei E. Gabrielian,[1] Weiniu Gan,[1] Wangmao Ge,[1] Fangcheng Gong,[1] Zhiping Gu,[1] Ping Guan,[1] Thomas J. Heiman,[1] Maureen E. Higgins,[1] Rui-Ru Ji,[1] Zhaoxi Ke,[1] Karen A. Ketchum,[1] Zhongwu Lai,[1] Yiding Lei,[1] Zhenya Li,[1] Jiayin Li,[1] Yong Liang,[1] Xiaoying Lin,[1] Fu Lu,[1] Gennady V. Merkulov,[1] Natalia Milshina,[1] Helen M. Moore,[1] Ashwinikumar K Naik,[1] Vaibhav A. Narayan,[1] Beena Neelam,[1] Deborah Nusskern,[1] Douglas B. Rusch,[1] Steven Salzberg,[12] Wei Shao,[1] Bixiong Shue,[1] Jingtao Sun,[1] Zhen Yuan Wang,[1] Aihui Wang,[1] Xin Wang,[1] Jian Wang,[1] Ming-Hui Wei,[1] Ron Wides,[13] Chunlin Xiao,[1] Chunhua Yan,[1] Alison Yao,[1] Jane Ye,[1] Ming Zhan,[1] Weiqing Zhang,[1] Hongyu Zhang,[1] Qi Zhao,[1] Liansheng Zheng,[1] Fei Zhong,[1] Wenyan Zhong,[1] Shiaoping C. Zhu,[1] Shaying Zhao,[12] Dennis Gilbert,[1] Suzanna Baumhueter,[1] Gene Spier,[1] Christine Carter,[1] Anibal Cravchik,[1] Trevor Woodage,[1] Feroze Ali,[1] Huijin An,[1] Aderonke Awe,[1] Danita Baldwin,[1] Holly Baden,[1] Mary Barnstead,[1] Ian Barrow,[1] Karen Beeson,[1] Dana Busam,[1] Amy Carver,[1] Angela Center,[1] Ming Lai Cheng,[1] Liz Curry,[1] Steve Danaher,[1] Lionel Davenport,[1] Raymond Desilets,[1] Susanne Dietz,[1] Kristina Dodson,[1] Lisa Doup,[1] Steven Ferriera,[1] Neha Garg,[1] Andres Gluecksmann,[1] Brit Hart,[1] Jason Haynes,[1] Charles Haynes,[1] Cheryl Heiner,[1] Suzanne Hladun,[1] Damon Hostin,[1] Jarrett Houck,[1] Timothy Howland,[1] Chinyere Ibegwam,[1] Jeffery Johnson,[1] Francis Kalush,[1] Lesley Kline,[1] Shashi Koduru,[1] Amy Love,[1] Felecia Mann,[1] David May,[1] Steven McCawley,[1] Tina McIntosh,[1] Ivy McMullen,[1] Mee Moy,[1] Linda Moy,[1] Brian Murphy,[1] Keith Nelson,[1] Cynthia Pfannkoch,[1] Eric Pratts,[1] Vinita Puri,[1] Hina Qureshi,[1] Matthew Reardon,[1] Robert Rodriguez,[1] Yu-Hui Rogers,[1] Deanna Romblad,[1] Bob Ruhfel,[1] Richard Scott,[1] Cynthia Sitter,[1] Michelle Smallwood,[1] Erin Stewart,[1] Renee Strong,[1] Ellen Suh,[1] Reginald Thomas,[1] Ni Ni Tint,[1] Sukyee Tse,[1] Claire Vech,[1] Gary Wang,[1] Jeremy Wetter,[1] Sherita Williams,[1] Monica Williams,[1] Sandra Windsor,[1] Emily Winn-Deen,[1] Keriellen Wolfe,[1] Jayshree Zaveri,[1] Karena Zaveri,[1] Josep F. Abril,[14] Roderic Guigó,[14] Michael J. Campbell,[1] Kimmen V. Sjolander,[1] Brian Karlak,[1] Anish Kejariwal,[1] Huaiyu Mi,[1] Betty Lazareva,[1] Thomas Hatton,[1] Apurva Narechania,[1] Karen Diemer,[1] Anushya Muruganujan,[1] Nan Guo,[1] Shinji Sato,[1] Vineet Bafna,[1] Sorin Istrail,[1] Ross Lippert,[1] Russell Schwartz,[1] Brian Walenz,[1] Shibu Yooseph,[1] David Allen,[1] Anand Basu,[1] James Baxendale,[1] Louis Blick,[1] Marcelo Caminha,[1] John Carnes-Stine,[1] Parris Caulk,[1] Yen-Hui Chiang,[1] My Coyne,[1] Carl Dahlke,[1] Anne Deslattes Mays,[1] Maria Dombroski,[1] Michael Donnelly,[1] Dale Ely,[1] Shiva Esparham,[1] Carl Fosler,[1] Harold Gire,[1] Stephen Glanowski,[1] Kenneth Glasser,[1] Anna Glodek,[1] Mark Gorokhov,[1] Ken Graham,[1] Barry Gropman,[1] Michael Harris,[1] Jeremy Heil,[1] Scott Henderson,[1] Jeffrey Hoover,[1] Donald Jennings,[1] Catherine Jordan,[1] James Jordan,[1] John Kasha,[1] Leonid Kagan,[1] Cheryl Kraft,[1] Alexander Levitsky,[1] Mark Lewis,[1] Xiangjun Liu,[1] John Lopez,[1] Daniel Ma,[1] William Majoros,[1] Joe McDaniel,[1] Sean Murphy,[1] Matthew Newman,[1] Trung Nguyen,[1] Ngoc Nguyen,[1] Marc Nodell,[1] Sue Pan,[1] Jim Peck,[1] Marshall Peterson,[1] William Rowe,[1] Robert Sanders,[1] John Scott,[1] Michael Simpson,[1] Thomas Smith,[1] Arlan Sprague,[1] Timothy Stockwell,[1] Russell Turner,[1] Eli Venter,[1] Mei Wang,[1] Meiyuan Wen,[1] David Wu,[1] Mitchell Wu,[1] Ashley Xia,[1] Ali Zandieh,[1] Xiaohong Zhu[1]

A 2.91-billion base pair (bp) consensus sequence of the euchromatic portion of the human genome was generated by the whole-genome shotgun sequencing method. The 14.8-billion bp DNA sequence was generated over 9 months from 27,271,853 high-quality sequence reads (5.11-fold coverage of the genome) from both ends of plasmid clones made from the DNA of five individuals. Two assembly strategies—a whole-genome assembly and a regional chromosome assembly—were used, each combining sequence data from Celera and the publicly funded genome effort. The public data were shredded into 550-bp segments to create a 2.9-fold coverage of those genome regions that had been sequenced, without including biases inherent in the cloning and assembly procedure used by the publicly funded group. This brought the effective coverage in the assemblies to eightfold, reducing the number and size of gaps in the final assembly over what would be obtained with 5.11-fold coverage. The two assembly strategies yielded very similar results that largely agree with independent mapping data. The assemblies effectively cover the euchromatic regions of the human chromosomes. More than 90% of the genome is in scaffold assemblies of 100,000 bp or more, and 25% of the genome is in scaffolds of 10 million bp or larger. Analysis of the genome sequence revealed 26,588 protein-encoding transcripts for which there was strong corroborating evidence and an additional ~12,000 computationally derived genes with mouse matches or other weak supporting evidence. Although gene-dense clusters are obvious, almost half the genes are dispersed in low G+C sequence separated by large tracts of apparently noncoding sequence. Only 1.1% of the genome is spanned by exons, whereas 24% is in introns, with 75% of the genome being intergenic DNA. Duplications of segmental blocks, ranging in size up to chromosomal lengths, are abundant throughout the genome and reveal a complex evolutionary history. Comparative genomic analysis indicates vertebrate expansions of genes associated with neuronal function, with tissue-specific developmental regulation, and with the hemostasis and immune systems. DNA sequence comparisons between the consensus sequence and publicly funded genome data provided locations of 2.1 million single-nucleotide polymorphisms (SNPs). A random pair of human haploid genomes differed at a rate of 1 bp per 1250 on average, but there was marked heterogeneity in the level of polymorphism across the genome. Less than 1% of all SNPs resulted in variation in proteins, but the task of determining which SNPs have functional consequences remains an open challenge.

Decoding of the DNA that constitutes the human genome has been widely anticipated for the contribution it will make toward understanding human evolution, the causation of disease, and the interplay between the environment and heredity in defining the human condition. A project with the goal of determining the complete nucleotide sequence of the human genome was first formally proposed in 1985 (*1*). In subsequent years, the idea met with mixed reactions in the scientific community (*2*). However, in 1990, the Human Genome Project (HGP) was officially initiated in the United States under the direction of the National Institutes of Health and the U.S. Department of Energy with a 15-year, $3 billion plan for completing the genome sequence. In 1998 we announced our intention to build a unique genome-sequencing facility, to determine the sequence of the human genome over a 3-year period. Here we report the penultimate milestone along the path toward that goal, a nearly complete sequence of the euchromatic portion of the human genome. The sequencing was performed by a whole-genome random shotgun method with subsequent assembly of the sequenced segments.

The modern history of DNA sequencing began in 1977, when Sanger reported his method for determining the order of nucleotides of DNA using chain-terminating nucleotide analogs (*3*). In the same year, the first human gene was isolated and sequenced (*4*). In 1986, Hood and co-workers (*5*) described an improvement in the Sanger sequencing method that included attaching fluorescent dyes to the nucleotides, which permitted them to be sequentially read by a computer. The first automated DNA sequencer, developed by Applied Biosystems in California in 1987, was shown to be successful when the sequences of two genes were obtained with this new technology (*6*). From early sequencing of human genomic regions (*7*), it became clear that cDNA sequences (which are reverse-transcribed from RNA) would be essential to annotate and validate gene predictions in the human genome. These studies were the basis in part for the development of the expressed sequence tag (EST) method of gene identification (*8*), which is a random selection, very high throughput sequencing approach to characterize cDNA libraries. The EST method led to the rapid discovery and mapping of human genes (*9*). The increasing numbers of human EST sequences necessitated the development of new computer algorithms to analyze large amounts of sequence data, and in 1993 at The Institute for Genomic Research (TIGR), an algorithm was developed that permitted assembly and analysis of hundreds of thousands of ESTs. This algorithm permitted characterization and annotation of human genes on the basis of 30,000 EST assemblies (*10*).

The complete 49-kbp bacteriophage lambda genome sequence was determined by a shotgun restriction digest method in 1982 (*11*). When considering methods for sequencing the smallpox virus genome in 1991 (*12*), a whole-genome shotgun sequencing method was discussed and subsequently rejected owing to the lack of appropriate software tools for genome assembly. However, in 1994, when a microbial genome-sequencing project was contemplated at TIGR, a whole-genome shotgun sequencing approach was considered possible with the TIGR EST assembly algorithm. In 1995, the 1.8-Mbp *Haemophilus influenzae* genome was completed by a whole-genome shotgun sequencing method (*13*). The experience with several subsequent genome-sequencing efforts established the broad applicability of this approach (*14*, *15*).

A key feature of the sequencing approach used for these megabase-size and larger genomes was the use of paired-end sequences (also called mate pairs), derived from subclone libraries with distinct insert sizes and cloning characteristics. Paired-end sequences are sequences 500 to 600 bp in length from both ends of double-stranded DNA clones of prescribed lengths. The success of using end sequences from long segments (18 to 20 kbp) of DNA cloned into bacteriophage lambda in assembly of the microbial genomes led to the suggestion (*16*) of an approach to simulta-

[1]Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA. [2]GenetixXpress, 78 Pacific Road, Palm Beach, Sydney 2108, Australia. [3]Berkeley *Drosophila* Genome Project, University of California, Berkeley, CA 94720, USA. [4]Department of Biology, Penn State University, 208 Mueller Lab, University Park, PA 16802, USA. [5]Department of Genetics, Case Western Reserve University School of Medicine, BRB-630, 10900 Euclid Avenue, Cleveland, OH 44106, USA. [6]Johns Hopkins University School of Medicine, Johns Hopkins Hospital, 600 North Wolfe Street, Blalock 1007, Baltimore, MD 21287–4922, USA. [7]Rockefeller University, 1230 York Avenue, New York, NY 10021–6399, USA. [8]New England BioLabs, 32 Tozer Road, Beverly, MA 01915, USA. [9]Division of Biology, 147-75, California Institute of Technology, 1200 East California Boulevard, Pasadena, CA 91125, USA. [10]Yale University School of Medicine, 333 Cedar Street, P.O. Box 208000, New Haven, CT 06520–8000, USA. [11]Applied Biosystems, 850 Lincoln Centre Drive, Foster City, CA 94404, USA. [12]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. [13]Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, 52900 Israel. [14]Grup de Recerca en Informàtica Mèdica, Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, 08003-Barcelona, Catalonia, Spain.

*To whom correspondence should be addressed. E-mail: humangenome@celera.com

neously map and sequence the human genome by means of end sequences from 150-kbp bacterial artificial chromosomes (BACs) (*17*, *18*). The end sequences spanned by known distances provide long-range continuity across the genome. A modification of the BAC end-sequencing (BES) method was applied successfully to complete chromosome 2 from the *Arabidopsis thaliana* genome (*19*).

In 1997, Weber and Myers (*20*) proposed whole-genome shotgun sequencing of the human genome. Their proposal was not well received (*21*). However, by early 1998, as less than 5% of the genome had been sequenced, it was clear that the rate of progress in human genome sequencing worldwide was very slow (*22*), and the prospects for finishing the genome by the 2005 goal were uncertain.

In early 1998, PE Biosystems (now Applied Biosystems) developed an automated, high-throughput capillary DNA sequencer, subsequently called the ABI PRISM 3700 DNA Analyzer. Discussions between PE Biosystems and TIGR scientists resulted in a plan to undertake the sequencing of the human genome with the 3700 DNA Analyzer and the whole-genome shotgun sequencing techniques developed at TIGR (*23*). Many of the principles of operation of a genome-sequencing facility were established in the TIGR facility (*24*). However, the facility envisioned for Celera would have a capacity roughly 50 times that of TIGR, and thus new developments were required for sample preparation and tracking and for whole-genome assembly. Some argued that the required 150-fold scale-up from the *H. influenzae* genome to the human genome with its complex repeat sequences was not feasible (*25*). The *Drosophila melanogaster* genome was thus chosen as a test case for whole-genome assembly on a large and complex eukaryotic genome. In collaboration with Gerald Rubin and the Berkeley *Drosophila* Genome Project, the nucleotide sequence of the 120-Mbp euchromatic portion of the *Drosophila* genome was determined over a 1-year period (*26*–*28*). The *Drosophila* genome-sequencing effort resulted in two key findings: (i) that the assembly algorithms could generate chromosome assemblies with highly accurate order and orientation with substantially less than 10-fold coverage, and (ii) that undertaking multiple interim assemblies in place of one comprehensive final assembly was not of value.

These findings, together with the dramatic changes in the public genome effort subsequent to the formation of Celera (*29*), led to a modified whole-genome shotgun sequencing approach to the human genome. We initially proposed to do 10-fold sequence coverage of the genome over a 3-year period and to make interim assembled sequence data available quarterly. The modifications included a plan to perform random shotgun sequencing to ~5-fold coverage and to use the unordered and unoriented BAC sequence fragments and subassemblies published in GenBank by the publicly funded genome effort (*30*) to accelerate the project. We also abandoned the quarterly announcements in the absence of interim assemblies to report.

Although this strategy provided a reasonable result very early that was consistent with a whole-genome shotgun assembly with eight-fold coverage, the human genome sequence is not as finished as the *Drosophila* genome was with an effective 13-fold coverage. However, it became clear that even with this reduced coverage strategy, Celera could generate an accurately ordered and oriented scaffold sequence of the human genome in less than 1 year. Human genome sequencing was initiated 8 September 1999 and completed 17 June 2000. The first assembly was completed 25 June 2000, and the assembly reported here was completed 1 October 2000. Here we describe the whole-genome random shotgun sequencing effort applied to the human genome. We developed two different assembly approaches for assembling the ~3 billion bp that make up the 23 pairs of chromosomes of the *Homo sapiens* genome. Any GenBank-derived data were shredded to remove potential bias to the final sequence from chimeric clones, foreign DNA contamination, or misassembled contigs. Insofar as a correctly and accurately assembled genome sequence with faithful order and orientation of contigs is essential for an accurate analysis of the human genetic code, we have devoted a considerable portion of this manuscript to the documentation of the quality of our reconstruction of the genome. We also describe our preliminary analysis of the human genetic code on the basis of computational methods. Figure 1 (see fold-out chart associated with this issue; files for each chromosome can be found in Web fig. 1 on *Science* Online at www.sciencemag.org/cgi/content/full/291/5507/1304/DC1) provides a graphical overview of the genome and the features encoded in it. The detailed manual curation and interpretation of the genome are just beginning.

To aid the reader in locating specific analytical sections, we have divided the paper into seven broad sections. A summary of the major results appears at the beginning of each section.

1 Sources of DNA and Sequencing Methods

Summary. This section discusses the rationale and ethical rules governing donor selection to ensure ethnic and gender diversity along with the methodologies for DNA extraction and library construction. The plasmid library construction is the first critical step in shotgun sequencing. If the DNA libraries are not uniform in size, nonchimeric, and do not randomly represent the genome, then the subsequent steps cannot accurately reconstruct the genome sequence. We used automated high-throughput DNA sequencing and the computational infrastructure to enable efficient tracking of enormous amounts of sequence information (27.3 million sequence reads; 14.9 billion bp of sequence). Sequencing and tracking from both ends of plasmid clones from 2-, 10-, and 50-kbp libraries were essential to the computational reconstruction of the genome. Our evidence indicates that the accurate pairing rate of end sequences was greater than 98%.

Various policies of the United States and the World Medical Association, specifically the Declaration of Helsinki, offer recommendations for conducting experiments with human subjects. We convened an Institutional Review Board (IRB) (*31*) that helped us establish the protocol for obtaining and using human DNA and the informed consent process used to enroll research volunteers for the DNA-sequencing studies reported here. We adopted several steps and procedures to protect the privacy rights and confidentiality of the research subjects (donors). These included a two-stage consent process, a secure random alphanumeric coding system for specimens and records, circumscribed contact with the subjects by researchers, and options for off-site contact of donors. In addition, Celera applied for and received a Certificate of Confidentiality from the Department of Health and Human Services. This Certificate authorized Celera to protect the privacy of the individuals who volunteered to be donors as provided in Section 301(d) of the Public Health Service Act 42 U.S.C. 241(d).

Celera and the IRB believed that the initial version of a completed human genome should be a composite derived from multiple donors of diverse ethnic backgrounds Prospective donors were asked, on a voluntary basis, to self-designate an ethnogeographic category (e.g., African-American, Chinese, Hispanic, Caucasian, etc.). We enrolled 21 donors (*32*).

Three basic items of information from each donor were recorded and linked by confidential code to the donated sample: age, sex, and self-designated ethnogeographic group. From females, ~130 ml of whole, heparinized blood was collected. From males, ~130 ml of whole, heparinized blood was

collected, as well as five specimens of se. , collected over a 6-week period. Permanent lymphoblastoid cell lines were created by Epstein-Barr virus immortalization. DNA from five subjects was selected for genomic DNA sequencing: two males and three females—one African-American, one Asian-Chinese, one Hispanic-Mexican, and two Caucasians (see Web fig. 2 on *Science* Online at www.sciencemag.org/cgi/content/291/5507/1304/DC1). The decision of whose DNA to sequence was based on a complex mix of factors, including the goal of achieving diversity as well as technical issues such as the quality of the DNA libraries and availability of immortalized cell lines.

1.1 Library construction and sequencing

Central to the whole-genome shotgun sequencing process is preparation of high-quality plasmid libraries in a variety of insert sizes so that pairs of sequence reads (mates) are obtained, one read from both ends of each plasmid insert. High-quality libraries have an equal representation of all parts of the genome, a small number of clones without inserts, and no contamination from such sources as the mitochondrial genome and *Escherichia coli* genomic DNA. DNA from each donor was used to construct plasmid libraries in one or more of three size classes: 2 kbp, 10 kbp, and 50 kbp (Table 1) (*33*).

In designing the DNA-sequencing process, we focused on developing a simple system that could be implemented in a robust and reproducible manner and monitored effectively (Fig. 2) (*34*).

Current sequencing protocols are based on the dideoxy sequencing method (*35*), which typically yields only 500 to 750 bp of sequence per reaction. This limitation on read length has made monumental gains in throughput a prerequisite for the analysis of large eukaryotic genomes. We accomplished this at the Celera facility, which occupies about 30,000 square feet of laboratory space and produces sequence data continuously at a rate of 175,000 total reads per day. The DNA-sequencing facility is supported by a high-performance computational facility (*36*).

The process for DNA sequencing was modular by design and automated. Intermodule sample backlogs allowed four principal modules to operate independently: (i) library transformation, plating, and colony picking; (ii) DNA template preparation; (iii) dideoxy sequencing reaction set-up and purification; and (iv) sequence determination with the ABI PRISM 3700 DNA Analyzer. Because the inputs and outputs of each module have been carefully matched and sample backlogs are continuously managed, sequencing has proceeded without a single day's interruption since the initiation of the *Drosophila* project in May 1999. The ABI 3700 is a fully automated capillary array sequencer and as such can be operated with a minimal amount of hands-on time, currently estimated at about 15 min per day. The capillary system also facilitates correct associations of sequencing traces with samples through the elimination of manual sample loading and lane-tracking errors associated with slab gels. About 65 production staff were hired and trained, and were rotated on a regular basis through the four production modules. A central laboratory information management system (LIMS) tracked all sample plates by unique bar code identifiers. The facility was supported by a quality control team that performed raw material and in-process testing and a quality assurance group with responsibilities including document control, validation, and auditing of the facility. Critical to the success of the scale-up was the validation of all software and instrumentation before implementation, and production-scale testing of any process changes.

1.2 Trace processing

An automated trace-processing pipeline has been developed to process each sequence file (*37*). After quality and vector trimming, the average trimmed sequence length was 543 bp, and the sequencing accuracy was exponentially distributed with a mean of 99.5% and with less than 1 in 1000 reads being less than 98% accurate (*26*). Each trimmed sequence was screened for matches to contaminants including sequences of vector alone, *E. coli* genomic DNA, and human mitochondrial DNA. The entire read for any sequence with a significant match to a contaminant was discarded. A total of 713 reads matched *E. coli* genomic DNA and 2114 reads matched the human mitochondrial genome.

1.3 Quality assessment and control

The importance of the base-pair level accuracy of the sequence data increases as the size and repetitive nature of the genome to be sequenced increases. Each sequence read must be placed uniquely in the ge-

Table 1. Celera-generated data input into assembly.

	Individual	Number of reads for different insert libraries				Total number of base pairs
		2 kbp	10 kbp	50 kbp	Total	
No. of sequencing reads	A	0	0	2,767,357	2,767,357	1,502,674,851
	B	11,736,757	7,467,755	66,930	19,271,442	10,464,393,006
	C	853,819	881,290	0	1,735,109	942,164,187
	D	952,523	1,046,815	0	1,999,338	1,085,640,534
	F	0	1,498,607	0	1,498,607	813,743,601
	Total	13,543,099	10,894,467	2,834,287	27,271,853	14,808,616,179
Fold sequence coverage (2.9-Gb genome)	A	0	0	0.52	0.52	
	B	2.20	1.40	0.01	3.61	
	C	0.16	1.17	0	0.32	
	D	0.18	0.20	0	0.37	
	F	0	0.28	0	0.28	
	Total	2.54	2.04	0.53	5.11	
Fold clone coverage	A	0	0	18.39	18.39	
	B	2.96	11.26	0.44	14.67	
	C	0.22	1.33	0	1.54	
	D	0.24	1.58	0	1.82	
	F	0	2.26	0	2.26	
	Total	3.42	16.43	18.84	38.68	
Insert size* (mean)	Average	1,951 bp	10,800 bp	50,715 bp		
Insert size* (SD)	Average	6.10%	8.10%	14.90%		
% Mates†	Average	74.50	80.80	75.60		

*Insert size and SD are calculated from assembly of mates on contigs. †% Mates is based on laboratory tracking of sequencing runs.

nome, and even a modest error rate can reduce the effectiveness of assembly. In addition, maintaining the validity of mate-pair information is absolutely critical for the algorithms described below. Procedural controls were established for maintaining the validity of sequence mate-pairs as sequencing reactions proceeded through the process, including strict rules built into the LIMS. The accuracy of sequence data produced by the Celera process was validated in the course of the *Drosophila* genome project (*26*). By collecting data for the entire human genome in a single facility, we were able to ensure uniform quality standards and the cost advantages associated with automation, an economy of scale, and process consistency.

2 Genome Assembly Strategy and Characterization

Summary. We describe in this section the two approaches that we used to assemble the genome. One method involves the computational combination of all sequence reads with shredded data from GenBank to generate an independent, nonbiased view of the genome. The second approach involves clustering all of the fragments to a region or chromosome on the basis of mapping information. The clustered data were then shredded and subjected to computational assembly. Both approaches provided essentially the same reconstruction of assembled DNA sequence with proper order and orientation. The second method provided slightly greater sequence coverage (fewer gaps) and was the principal sequence used for the analysis phase. In addition, we document the completeness and correctness of this assembly process

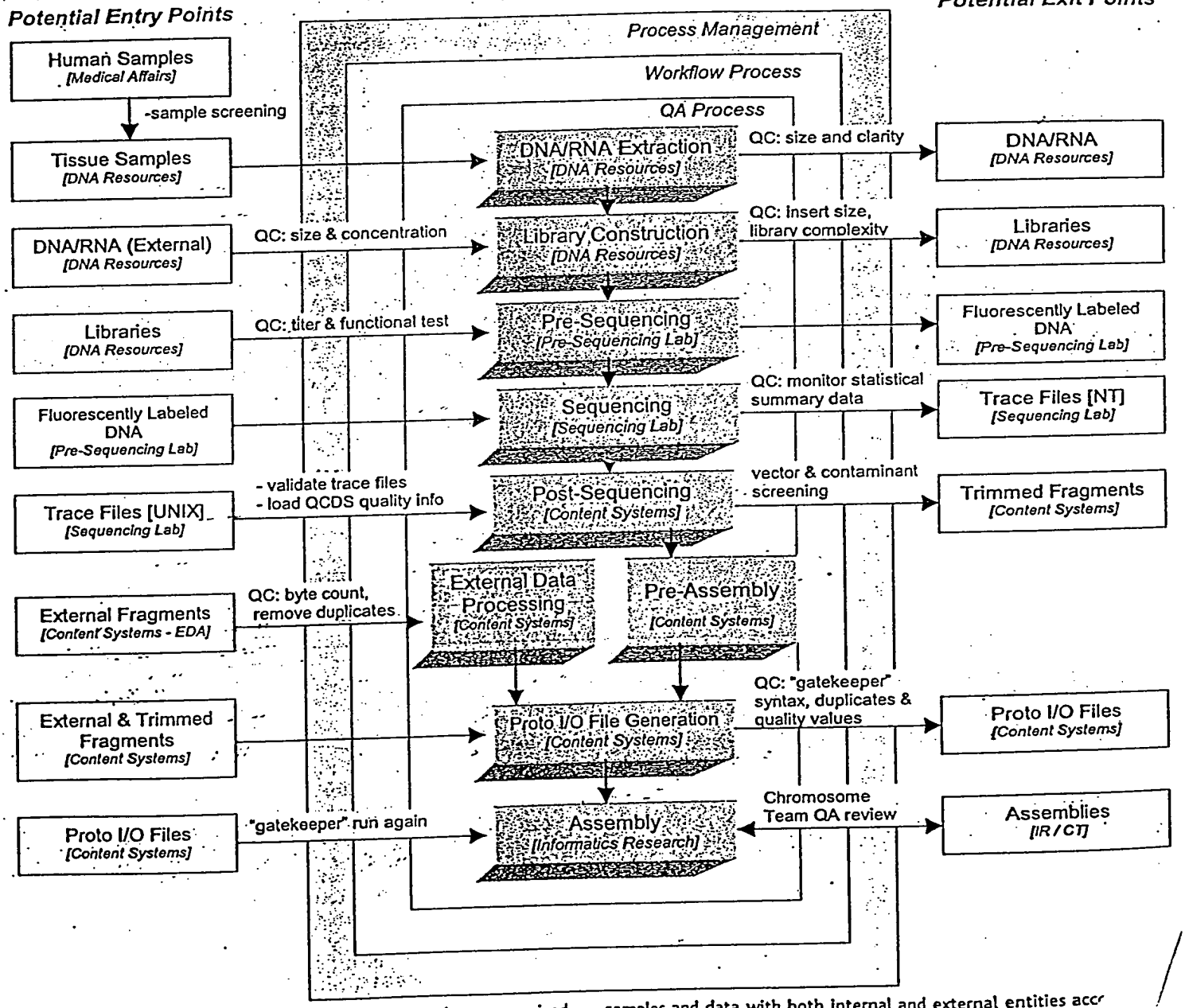


Fig. 2. Flow diagram for sequencing pipeline. Samples are received, selected, and processed in compliance with standard operating procedures, with a focus on quality within and across departments. Each process has defined inputs and outputs with the capability to exchange samples and data with both internal and external entities acc defined quality guidelines. Manufacturing pipeline processes quality control measures, and responsible parties are indic described further in the text.

and provide a comparison to the public genome sequence, which was reconstructed largely by an independent BAC-by-BAC approach. Our assemblies effectively covered the euchromatic regions of the human chromosomes. More than 90% of the genome was in scaffold assemblies of 100,000 bp or greater, and 25% of the genome was in scaffolds of 10 million bp or larger.

Shotgun sequence assembly is a classic example of an inverse problem: given a set of reads randomly sampled from a target sequence, reconstruct the order and the position of those reads in the target. Genome assembly algorithms developed for *Drosophila* have now been extended to assemble the ~25-fold larger human genome. Celera assemblies consist of a set of contigs that are ordered and oriented into scaffolds that are then mapped to chromosomal locations by using known markers. The contigs consist of a collection of overlapping sequence reads that provide a consensus reconstruction for a contiguous interval of the genome. Mate pairs are a central component of the assembly strategy. They are used to produce scaffolds in which the size of gaps between consecutive contigs is known with reasonable precision. This is accomplished by observing that a pair of reads, one of which is in one contig, and the other of which is in another, implies an orientation and distance between the two contigs (Fig. 3). Finally, our assemblies did not incorporate all reads into the final set of reported scaffolds. This set of unincorporated reads is termed "chaff," and typically consisted of reads from within highly repetitive regions, data from other organisms introduced through various routes as found in many genome projects, and data of poor quality or with untrimmed vector.

2.1 Assembly data sets

We used two independent sets of data for our assemblies. The first was a random shotgun data set of 27.27 million reads of average length 543 bp produced at Celera. This consisted largely of mate-pair reads from 16 libraries constructed from DNA samples taken from five different donors. Libraries with insert sizes of 2, 10, and 50 kbp were used. By looking at how mate pairs from a library were positioned in known sequenced stretches of the genome, we were able to characterize the range of insert sizes in each library and determine a mean and standard deviation. Table 1 details the number of reads, sequencing coverage, and clone coverage achieved by the data set. The clone coverage is the coverage of the genome in cloned DNA, considering the entire insert of each clone that has sequence from both ends. The clone coverage provides a measure of the amount of physical DNA coverage of the genome. Assuming a genome size of 2.9 Gbp, the Celera trimmed sequences gave a 5.1× coverage of the genome, and clone coverage was 3.42×, 16.40×, and 18.84× for the 2-, 10-, and 50-kbp libraries, respectively, for a total of 38.7× clone coverage.

The second data set was from the publicly funded Human Genome Project (PFP) and is primarily derived from BAC clones (*30*). The BAC data input to the assemblies came from a download of GenBank on 1 September 2000 (Table 2) totaling 4443.3 Mbp of sequence. The data for each BAC is deposited at one of four levels of completion. Phase 0 data are a set of generally unassembled sequencing reads from a very light shotgun of the BAC, typically less than 1×. Phase 1 data are unordered assemblies of contigs, which we call BAC contigs or bactigs. Phase 2 data are ordered assemblies of bactigs. Phase 3 data are complete BAC sequences. In the past 2 years the PFP has focused on a product of lower quality and completeness, but on a faster time-course, by concentrating on the production of Phase 1 data from a 3× to 4× light-shotgun of each BAC clone.

We screened the bactig sequences for contaminants by using the BLAST algorithm against three data sets: (i) vector sequences in Univec core (*38*), filtered for a 25-bp match at 98% sequence identity at the ends of the sequence and a 30-bp match internal to the sequence; (ii) the nonhuman portion of the High Throughput Genomic (HTG) Seqences division of GenBank (*39*), filtered at 200 bp at 98%; and (iii) the nonredundant nucleotide sequences from GenBank without primate and human virus entries, filtered at 200 bp at 98%. Whenever 25 bp or more of vector was found within 50 bp of the end of a contig, the tip up to the matching vector was excised. Under these criteria we removed 2.6 Mbp of possible contaminant and vector from the Phase 3 data, 61.0 Mbp from the Phase 1 and 2 data, and 16.1 Mbp from the Phase 0 data (Table 2). This left us with a total of 4363.7 Mbp of PFP sequence data 20% finished, 75% rough-draft (Phase 1 and 2), and 5% single sequencing reads (Phase 0). An additional 104,018 BAC end-sequence mate pairs were also downloaded and included in the data sets for both assembly processes (*18*).

2.2 Assembly strategies

Two different approaches to assembly were pursued. The first was a whole-genome assembly process that used Celera data and the PFP data in the form of additional synthetic shotgun data, and the second was a compartmentalized assembly process that first partitioned the Celera and PFP data into sets localized to large chromosomal segments and then performed ab initio shotgun assembly on each set. Figure 4 gives a schematic of the overall process flow.

For the whole-genome assembly, the PFP data was first disassembled or "shredded" into a synthetic shotgun data set of 550-bp reads that form a perfect 2× covering of the bactigs. This resulted in 16.05 million "faux" reads that were sufficient to cover the genome 2.96× because of redundancy in the BAC data set, without incorporating the biases inherent in the PFP assembly process. The combined data set of 43.32 million reads (8×), and all associated mate-pair information, were then subjected to our whole-genome assembly algorithm to produce a reconstruction of the genome. Neither the location of a BAC in the genome nor its assembly of bactigs was used in this process. Bactigs were shredded into reads because we found strong evidence that 2.13% of them were misassembled (*40*). Furthermore, BAC location

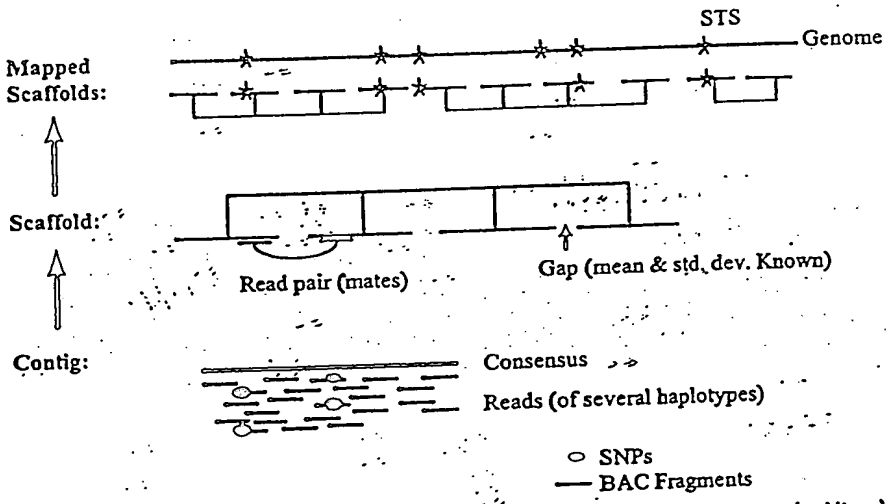


Fig. 3. Anatomy of whole-genome assembly. Overlapping shredded bactig fragments (red lines) and internally derived reads from five different individuals (black lines) are combined to produce a contig and a consensus sequence (green line). Contigs are connected into scaffolds (red) by using mate pair information. Scaffolds are then mapped to the genome (gray line) with STS (blue star) physical map information.

information was ignored because some BACs were not correctly placed on the PFP physical map and because we found strong evidence that at least 2.2% of the BACs contained sequence data that were not part of the given BAC (*41*), possibly as a result of sample-tracking errors (see below). In short, we performed a true, ab initio whole-genome assembly in which we took the expedient of deriving additional sequence coverage, but not mate pairs, assembled bactigs, or genome locality, from some externally generated data.

Table 2. GenBank data input into assembly.

Center	Statistics	Completion phase sequence		
		0	1 and 2	3
Whitehead Institute/ MIT Center for Genome Research, USA	Number of accession records	2,825	6,533	363
	Number of contigs	243,786	138,023	363
	Total base pairs	194,490,158	1,083,848,245	48,829,358
	Total vector masked (bp)	1,553,597	875,618	2,202
	Total contaminant masked (bp)	13,654,482	4,417,055	98,028
	Average contig length (bp)	798	7,853	134,516
Washington University, USA	Number of accession records	19	3,232	1,300
	Number of contigs	2,127	61,812	1,300
	Total base pairs	1,195,732	561,171,788	164,214,395
	Total vector masked (bp)	21,604	270,942	8,287
	Total contaminant masked (bp)	22,469	1,476,141	469,487
	Average contig length (bp)	562	9,079	126,319
Baylor College of Medicine, USA	Number of accession records	0	1,626	363
	Number of contigs	0	44,861	363
	Total base pairs	0	265,547,066	49,017,104
	Total vector masked (bp)	0	218,769	4,960
	Total contaminant masked (bp)	0	1,784,700	485,137
	Average contig length (bp)	0	5,919	135,033
Production Sequencing Facility, DOE Joint Genome Institute, USA	Number of accession records	135	2,043	754
	Number of contigs	7,052	34,938	754
	Total base pairs	8,680,214	294,249,631	60,975,328
	Total vector masked (bp)	22,644	162,651	7,274
	Total contaminant masked (bp)	665,818	4,642,372	118,387
	Average contig length (bp)	1,231	8,422	80,867
The Institute of Physical and Chemical Research (RIKEN), Japan	Number of accession records	0	1,149	300
	Number of contigs	0	25,772	300
	Total base pairs	0	182,812,275	20,093,926
	Total vector masked (bp)	0	203,792	2,371
	Total contaminant masked (bp)	0	308,426	27,781
	Average contig length (bp)	0	7,093	66,978
Sanger Centre, UK	Number of accession records	0	4,538	2,599
	Number of contigs	0	74,324	2,599
	Total base pairs	0	689,059,692	246,118,000
	Total vector masked (bp)	0	427,326	25,054
	Total contaminant masked (bp)	0	2,066,305	374,561
	Average contig length (bp)	0	9,271	94,697
Others*	Number of accession records	42	1,894	3,458
	Number of contigs	5,978	29,898	3,458
	Total base pairs	5,564,879	283,358,877	246,474,157
	Total vector masked (bp)	57,448	279,477	32,136
	Total contaminant masked (bp)	575,366	1,616,665	1,791,849
	Average contig length (bp)	931	9,478	71,277
All centers combined†	Number of accession records	3,021	21,015	9,137
	Number of contigs	258,943	409,628	9,137
	Total base pairs	209,930,983	3,360,047,574	835,722,268
	Total vector masked (bp)	1,655,293	2,438,575	82,284
	Total contaminant masked (bp)	14,918,135	16,311,664	3,365,230
	Average contig length (bp)	811	8,203	91,466

*Other centers contributing at least 0.1% of the sequence include: Chinese National Human Genome Center; Genomanalyse Gesellschaft fuer Biotechnologische Forschung mbH; Genome Therapeutics Corporation; GENOSCOPE; Chinese Academy of Sciences; Institute of Molecular Biotechnology; Keio University School of Medicine; Lawrence Livermore National Laboratory; Cold Spring Harbor Laboratory; Los Alamos National Laboratory; Max-Planck Institut fuer Molekulare, Genetik; Japan Science and Technology Corporation; Stanford University; The Institute for Genomic Research; The Institute of Physical and Chemical Research, Gene Bank; The University of Oklahoma; University of Texas Southwestern Medical Center, University of Washington. †The 4,405,700,825 bases contributed by all centers were shredded into faux reads resulting in 2.96× coverage of the genome.

In the compartmentalized shotgun assembly (CSA), Celera and PFP data were partitioned into the largest possible chromosomal segments or "components" that could be determined with confidence, and then shotgun assembly was applied to each partitioned subset wherein the bactig data were again shredded into faux reads to ensure an independent ab initio assembly of the component. By subsetting the data in this way, the overall computational effort was reduced and the effect of interchromosomal duplications was ameliorated. This also resulted in a reconstruction of the genome that was relatively independent of the whole-genome assembly results so that the two assemblies could be compared for consistency. The quality of the partitioning into components was crucial so that different genome regions were not mixed together. We constructed components from (i) the longest scaffolds of the sequence from each BAC and (ii) assembled scaffolds of data unique to Celera's data set. The BAC assemblies were obtained by a combining assembler that used the bactigs and the 5× Celera data mapped to those bactigs as input. This effort was undertaken as an interim step solely because the more accurate and complete the scaffold for a given sequence stretch, the more accurately one can tile these scaffolds into contiguous components on the basis of sequence overlap and mate-pair information. We further visually inspected and edited the scaffold tiling of the components to further increase its accuracy. For the final CSA assembly, all but the partitioning was ignored and an independent, ab initio reconstruction of the sequence in each component was obtained by applying our whole-genome assembly algorithm to the partitioned, relevant Celera data and the shredded, faux reads of the partitioned, relevant bactig data.

2.3 Whole-genome assembly

The algorithms used for whole-genome assembly (WGA) of the human genome were enhancements to those used to produce the sequence of the *Drosophila* genome reported in detail in (*28*).

The WGA assembler consists of a pipeline composed of five principal stages: Screener, Overlapper, Unitigger, Scaffolder, and Repeat Resolver, respectively. The Screener finds and marks all microsatellite repeats with less than a 6-bp element, and screens out other known interspersed repeat elements, including Alu, Line, and ribosomal DNA. Marked regions get searched for overlaps, while screened regions do not get searched and must be part of an overlap that involves two matching segments.

The Overlapper compares every read against every other read in search of complete end-to-end overlaps of at least 40 bp and with no more than 6% differences in the match. Because all data are scrupulously vector-trimmed, the Overlapper can insist on complete overlap matches. Computing the set of all overlaps took roughly 10,000 CPU hours with a suite of four-processor Alpha SMPs with 4 gigabytes of RAM. This took 4 to 5 days in elapsed time with 40 such machines operating in parallel.

Every overlap computed above is statistically a 1-in-10^{17} event and thus not a coincidental event. What makes assembly combinatorially difficult is that while many overlaps are actually sampled from overlapping regions of the genome, and thus imply that the sequence reads should be assembled together, even more overlaps are actually from two distinct copies of a low-copy repeated element not screened above, thus constituting an error if put together. We call the former "true overlaps" and the latter "repeat-induced overlaps." The assembler must avoid choosing repeat-induced overlaps, especially early in the process.

We achieve this objective in the Unitigger. We first find all assemblies of reads that appear to be uncontested with respect to all other reads. We call the contigs formed from these subassemblies unitigs (for uniquely assembled contigs). Formally, these unitigs are the uncontested interval subgraphs of the graph of all overlaps (*42*). Unfortunately, although empirically many of these assemblies are correct (and thus involve only true overlaps), some are in fact collections of reads from several copies of a repetitive element that have been overcollapsed into a single subassembly. However, the overcollapsed unitigs are easily identified because their average coverage depth is too high to be consistent with the overall level of sequence coverage. We developed a simple statistical discriminator that gives the logarithm of the odds ratio that a unitig is composed of unique DNA or of a repeat consisting of two or more copies. The discriminator, set to a sufficiently stringent threshold, identifies a subset of the unitigs that we are certain are correct. In addition, a second, less stringent threshold identifies a subset of remaining unitigs very likely to be correctly assembled, of which we select those that will consistently scaffold (see below), and thus are again almost certain to be correct. We call the union of these two sets U-unitigs. Empirically, we found from a 6× simulated shotgun of human chromosome 22 that we get U-unitigs covering 98% of the stretches of unique DNA that are >2 kbp long. We are further able to identify the boundary of the start of a repetitive element at the ends of a U-unitig and leverage this so that U-unitigs span more than 93% of all singly interspersed Alu elements and other 100-to 400-bp repetitive segments.

The result of running the Unitigger was thus a set of correctly assembled subcontigs covering an estimated 73.6% of the human genome. The Scaffolder then proceeded to use mate-pair information to link these together into scaffolds. When there are two or more mate pairs that imply that a given pair of U-unitigs are at a certain distance and orientation with respect to each other, the probability of this being wrong is again roughly 1 in 10^{10}, assuming that mate pairs are false less than 2% of the time. Thus, one can with high confidence link together all U-unitigs that are linked by at least two 2- or 10-kbp mate pairs producing intermediate-sized scaffolds that are then recursively linked together by confirming 50-kbp mate pairs and BAC end sequences. This process yielded scaffolds that are on the order of megabase pairs in size with gaps between their contigs that generally correspond to repetitive elements and occasionally to small sequencing gaps. These scaffolds reconstruct the majority of the unique sequence within a genome.

For the *Drosophila* assembly, we engaged in a three-stage repeat resolution strategy where each stage was progressively more aggressive and thus more likely to make a mistake. For the human assembly, we continued to use the first "Rocks" substage where all unitigs with a good, but not definitive, discriminator score are placed in a scaffold gap. This was done with the condition that two or more mate pairs with one of their reads already in the scaffold unambiguously place the unitig in the given gap. We estimate the probability of inserting a unitig into an incorrect gap with this strategy to be less than 10^{-7} based on a probabilistic analysis.

We revised the ensuing "Stones" substage of the human assembly, making it more like the mechanism suggested in our earlier work (*43*). For each gap, every read R that is placed in the gap by virtue of its mated pair M being in a contig of the scaffold and implying R's placement is collected. Celera's mate-pairing information is correct more than 99% of the time. Thus, almost every, but not all, of the reads in the set belong in the gap, and when a read does not belong it rarely agrees with the remainder of the reads. Therefore, we simply assemble this set of reads within the gap, eliminating any reads that conflict with the assembly. This operation proved much more reliable than the one it replaced for the *Drosophila* assembly; in the assembly of a simulated shotgun data set of human chromo-

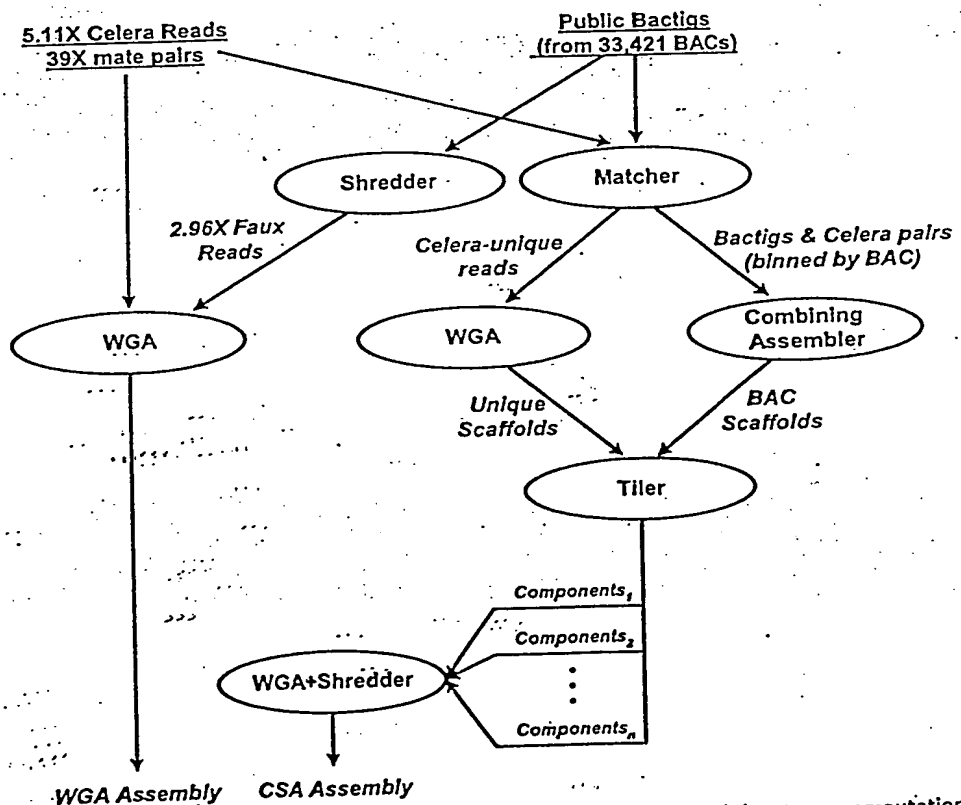


Fig. 4. Architecture of Celera's two-pronged assembly strategy. Each oval denotes a computation process performing the function indicated by its label, with the labels on arcs between ovals describing the nature of the objects produced and/or consumed by a process. This figure summarizes the discussion in the text that defines the terms and phrases used.

some 22, all stones were placed correctly.

The final method of resolving gaps is to fill them with assembled BAC data that cover the gap. We call this external gap "walking." We did not include the very aggressive "Pebbles" substage described in our *Drosophila* work, which made enough mistakes so as to produce repeat reconstructions for long interspersed elements whose quality was only 99.62% correct. We decided that for the human genome it was philosophically better not to introduce a step that was certain to produce less than 99.99% accuracy. The cost was a somewhat larger number of gaps of somewhat larger size.

At the final stage of the assembly process, and also at several intermediate points, a consensus sequence of every contig is produced. Our algorithm is driven by the principle of maximum parsimony, with quality-value–weighted measures for evaluating each base. The net effect is a Bayesian estimate of the correct base to report at each position. Consensus generation uses Celera data whenever it is present. In the event that no Celera data cover a given region, the BAC data sequence is used.

A key element of achieving a WGA of the human genome was to parallelize the Overlapper and the central consensus sequence–constructing subroutines. In addition, memory was a real issue—a straightforward application of the software we had built for *Drosophila* would have required a computer with a 600-gigabyte RAM. By making the Overlapper and Unitigger incremental, we were able to achieve the same computation with a maximum of instantaneous usage of 28 gigabytes of RAM. Moreover, the incremental nature of the first three stages allowed us to continually update the state of this part of the computation as data were delivered and then perform a 7-day run to complete Scaffolding and Repeat Resolution whenever desired. For our assembly operations, the total compute infrastructure consists of 10 four-processor SMPs with 4 gigabytes of memory per cluster (Compaq's ES40, Regatta) and a 16-processor NUMA machine with 64 gigabytes of memory (Compaq's GS160, Wildfire). The total compute for a run of the assembler was roughly 20,000 CPU hours.

The assembly of Celera's data, together with the shredded bactig data, produced a set of scaffolds totaling 2.848 Gbp in span and consisting of 2.586 Gbp of sequence. The chaff, or set of reads not incorporated in the assembly, numbered 11.27 million (26%), which is consistent with our experience for *Drosophila*. More than 84% of the genome was covered by scaffolds >100 kbp long, and these averaged 91% sequence and 9% gaps with a total of 2.297 Gbp of sequence. There were a total of 93,857 gaps among the 1637 scaffolds >100 kbp. The average scaffold size was 1.5 Mbp, the average contig size was 24.06 kbp, and the average gap size was 2.43 kbp, where the distribution of each was essentially exponential. More than 50% of all gaps were less than 50(bp long, >62% of all gaps were less than 1 kbp long, and no gap was >100 kbp long. Similarly, more than 65% of the sequence is in contigs >30 kbp, more than 31% is in contigs >100 kbp, and the largest contig was 1.22 Mbp long. Table 3 gives detailed summary statistics for the structure of this assembly with a direct comparison to the compartmentalized shotgun assembly.

2.4 Compartmentalized shotgun assembly

In addition to the WGA approach, we pursued a localized assembly approach that was intended to subdivide the genome into segments, each of which could be shotgun assembled individually. We expected that this would help in resolution of large interchromosomal duplications and improve the statistics for calculating U-unitigs. The compartmentalized assembly process involved clustering Celera reads and bactigs into large, multiple megabase regions of the genome, and then running the WGA assembler on the Celera data and shredded, faux reads obtained from the bactig data.

The first phase of the CSA strategy was to separate Celera reads into those that matched the BAC contigs for a particular PFP BAC entry, and those that did not match any public data. Such matches must be guaranteed to

Table 3. Scaffold statistics for whole-genome and compartmentalized shotgun assemblies.

	Scaffold size				
	All	>30 kbp	>100 kbp	>500 kbp	>1000 kbp
	Compartmentalized shotgun assembly				
No. of bp in scaffolds (including intrascaffold gaps)	2,905,568,203	2,748,892,430	2,700,489,906	2,489,357,260	2,248,689,128
No. of bp in contigs	2,653,979,733	2,524,251,302	2,491,538,372	2,320,648,201	2,106,521,902
No. of scaffolds	53,591	2,845	1,935	1,060	721
No. of contigs	170,033	112,207	107,199	93,138	82,009
No. of gaps	116,442	109,362	105,264	92,078	81,288
No. of gaps ≤1 kbp	72,091	69,175	67,289	59,915	53,354
Average scaffold size (bp)	54,217	966,219	1,395,602	2,348,450	3,118,848
Average contig size (bp)	15,609	22,496	23,242	24,916	25,686
Average intrascaffold gap size (bp)	2,161	2,054	1,985	1,832	1,749
Largest contig (bp)	1,988,321	1,988,321	1,988,321	1,988,321	1,988,321
% of total contigs	100	95	94	87	79
	Whole-genome assembly				
No. of bp in scaffolds (including intrascaffold gaps)	2,847,890,390	2,574,792,618	2,525,334,447	2,328,535,466	2,140,943,032
No. of bp in contigs	2,586,634,108	2,334,343,339	2,297,678,935	2,143,002,184	1,983,305,432
No. of scaffolds	118,968	2,507	1,637	818	554
No. of contigs	221,036	99,189	95,494	84,641	76,285
No. of gaps	102,068	96,682	93,857	83,823	75,731
No. of gaps ≤1 kbp	62,356	60,343	59,156	54,079	49,592
Average scaffold size (bp)	23,938	1,027,041	1,542,660	2,846,620	3,864,518
Average contig size (bp)	11,702	23,534	24,061	25,319	25,999
Average intrascaffold gap size (bp)	2,560	2,487	2,426	2,213	2,082
Largest contig (bp)	1,224,073	1,224,073	1,224,073	1,224,073	1,224,073
% of total contigs	100	90	89	83	77

properly place a Celera read, so all reads were first masked against a library of common repetitive elements, and only matches of at least 40 bp to unmasked portions of the read constituted a hit. Of Celera's 27.27 million reads, 20.76 million matched a bactig and another 0.62 million reads, which did not have any matches, were nonetheless identified as belonging in the region of the bactig's BAC because their mate matched the bactig. Of the remaining reads, 2.92 million were completely screened out and so could not be matched, but the other 2.97 million reads had unmasked sequence totaling 1.189 Gbp that were not found in the GenBank data set. Because the Celera data are 5.11× redundant, we estimate that 240 Mbp of unique Celera sequence is not in the GenBank data set.

In the next step of the CSA process, a combining assembler took the relevant 5× Celera reads and bactigs for a BAC entry, and produced an assembly of the combined data for that locale. These high-quality sequence reconstructions were a transient result whose utility was simply to provide more reliable information for the purposes of their tiling into sets of overlapping and adjacent scaffold sequences in the next step. In outline, the combining assembler first examines the set of matching Celera reads to determine if there are excessive pileups indicative of unscreened repetitive elements. Wherever these occur, reads in the repeat region whose mates have not been mapped to consistent positions are removed. Then all sets of mate pairs that consistently imply the same relative position of two bactigs are bundled into a link and weighted according to the number of mates in the bundle. A "greedy" strategy then attempts to order the bactigs by selecting bundles of mate-pairs in order of their weight. A selected mate-pair bundle can tie together two formative scaffolds. It is incorporated to form a single scaffold only if it is consistent with the majority of links between contigs of the scaffold. Once scaffolding is complete, gaps are filled by the "Stones" strategy described above for the WGA assembler.

The GenBank data for the Phase 1 and 2 BACs consisted of an average of 19.8 bactigs per BAC of average size 8099 bp. Application of the combining assembler resulted in individual Celera BAC assemblies being put together into an average of 1.83 scaffolds (median of 1 scaffold) consisting of an average of 8.57 contigs of average size 18,973 bp. In addition to defining order and orientation of the sequence fragments, there were 57% fewer gaps in the combined result. For Phase 0 data, the average GenBank entry consisted of 91.52 reads of average length 784 bp. Application of the combining assembler resulted in an average of 54.8 scaffolds consisting of an average of 58.1 contigs of average size 873 bp. Basically, some small amount of assembly took place, but not enough Celera data were matched to truly assemble the 0.5× to 1× data set represented by the typical Phase 0 BACs. The combining assembler was also applied to the Phase 3 BACs for SNP identification, confirmation of assembly, and localization of the Celera reads. The phase 0 data suggest that a combined whole-genome shotgun data set and 1× light-shotgun of BACs will not yield good assembly of BAC regions; at least 3× light-shotgun of each BAC is needed.

The 5.89 million Celera fragments not matching the GenBank data were assembled with our whole-genome assembler. The assembly resulted in a set of scaffolds totaling 442 Mbp in span and consisting of 326 Mbp of sequence. More than 20% of the scaffolds were >5 kbp long, and these averaged 63% sequence and 27% gaps with a total of 302 Mbp of sequence. All scaffolds >5 kbp were forwarded along with all scaffolds produced by the combining assembler to the subsequent tiling phase.

At this stage, we typically had one or two scaffolds for every BAC region constituting at least 95% of the relevant sequence, and a collection of disjoint Celera-unique scaffolds. The next step in developing the genome components was to determine the order and overlap tiling of these BAC and Celera-unique scaffolds across the genome. For this, we used Celera's 50-kbp mate-pairs information, and BAC-end pairs (*18*) and sequence tagged site (STS) markers (*44*) to provide long-range guidance and chromosome separation. Given the relatively manageable number of scaffolds, we chose not to produce this tiling in a fully automated manner, but to compute an initial tiling with a good heuristic and then use human curators to resolve discrepancies or missed join opportunities. To this end, we developed a graphical user interface that displayed the graph of tiling overlaps and the evidence for each. A human curator could then explore the implication of mapped STS data, dot-plots of sequence overlap, and a visual display of the mate-pair evidence supporting a given choice. The result of this process was a collection of "components," where each component was a tiled set of BAC and Celera-unique scaffolds that had been curator-approved. The process resulted in 3845 components with an estimated span of 2.922 Gbp.

In order to generate the final CSA, we assembled each component with the WGA algorithm. As was done in the WGA process, the bactig data were shredded into a synthetic 2× shotgun data set in order to give the assembler the freedom to independently assemble the data. By using faux reads rather than bactigs, the assembly algorithm could correct errors in the assembly of bactigs and remove chimeric content in a PFP data entry. C. ...eric or contaminating sequence (from another part of the genome) would not be incorporated into the reassembly of the component because it did not belong there. In effect, the previous steps in the CSA process served only to bring together Celera fragments and PFP data relevant to a large contiguous segment of the genome, wherein we applied the assembler used for WGA to produce an ab initio assembly of the region.

WGA assembly of the components resulted in a set of scaffolds totaling 2.906 Gbp in span and consisting of 2.654 Gbp of sequence. The chaff, or set of reads not incorporated into the assembly, numbered 6.17 million, or 22%. More than 90.0% of the genome was covered by scaffolds spanning >100 kbp long, and these averaged 92.2% sequence and 7.8% gaps with a total of 2.492 Gbp of sequence. There were a total of 105,264 gaps among the 107,199 contigs that belong to the 1940 scaffolds spanning >100 kbp. The average scaffold size was 1.4 Mbp, the average contig size was 23.24 kbp, and the average gap size was 2.0 kbp where each distribution of sizes was exponential. As such, averages tend to be underrepresentative of the majority of the data. Figure 5 shows a histogram of the bases in scaffolds of various size ranges. Consider also that more than 49% of all gaps were <500 bp long, more than 62% of all gaps were <1 kbp, and all gaps are <100 kbp long. Similarly, more than 73% of the sequence is in contigs > 30 kbp, more than 49% is in contigs >100 kbp, and the largest contig was 1.99 Mbp long. Table 3 provides summary statistics for the structure of this assembly with a direct comparison to the WGA assembly.

2.5 Comparison of the WGA and CSA scaffolds

Having obtained two assemblies of the human genome via independent computational processes (WGA and CSA), we compared scaffolds from the two assemblies as another means of investigating their completeness, consistency, and contiguity. From each assembly, a set of reference scaffolds containing at least 1000 fragments (Celera sequencing reads or bactig shreds) was obtained; this amounted to 2218 WGA scaffolds and 1717 CSA scaffolds, for a total of 2.087 Gbp and 2.474 Gbp. The sequence of each reference scaffold was compared to the sequence of all scaffolds from the other assembly with which it shared at least 20 fragments or at least 20% of the fragments of the smaller scaffold. For each such comparison, all matches of at least 200 bp with at most 2% mismatch were tabulated.

From this tabulation, we estimated the amount of unique sequence in each assembly in two ways. The first was to determine the number of bases of each assembly that were

not covered by a matching segment in the other assembly. Some 82.5 Mbp of the WGA (3.95%) was not covered by the CSA, whereas 204.5 Mbp (8.26%) of the CSA was not covered by the WGA. This estimate did not require any consistency of the assemblies or any uniqueness of the matching segments. Thus, another analysis was conducted in which matches of less than 1 kbp between a pair of scaffolds were excluded unless they were confirmed by other matches having a consistent order and orientation. This gives some measure of consistent coverage: 1.982 Gbp (95.00%) of the WGA is covered by the CSA, and 2.169 Gbp (87.69%) of the CSA is covered by the WGA by this more stringent measure.

The comparison of WGA to CSA also permitted evaluation of scaffolds for structural inconsistencies. We looked for instances in which a large section of a scaffold from one assembly matched only one scaffold from the other assembly, but failed to match over the full length of the overlap implied by the matching segments. An initial set of candidates was identified automatically, and then each candidate was inspected by hand. From this process, we identified 31 instances in which the assemblies appear to disagree in a nonlocal fashion. These cases are being further evaluated to determine which assembly is in error and why.

In addition, we evaluated local inconsistencies of order or orientation. The following results exclude cases in which one contig in one assembly corresponds to more than one overlapping contig in the other assembly (as long as the order and orientation of the latter agrees with the positions they match in the former). Most of these small rearrangements involved segments on the order of hundreds of base pairs and rarely >1 kbp. We found a total of 295 kbp (0.012%) in the CSA assemblies that were locally inconsistent with the WGA assemblies, whereas 2.108 Mbp (0.11%) in the WGA assembly were inconsistent with the CSA assembly.

The CSA assembly was a few percentage points better in terms of coverage and slightly more consistent than the WGA, because it was in effect performing a few thousand shotgun assemblies of megabase-sized problems, whereas the WGA is performing a shotgun assembly of a gigabase-sized problem. When one considers the increase of two-and-a-half orders of magnitude in problem size, the information loss between the two is remarkably small. Because CSA was logistically easier to deliver and the better of the two results available at the time when downstream analyses needed to be begun, all subsequent analysis was performed on this assembly.

2.6 Mapping scaffolds to the genome

The final step in assembling the genome was to order and orient the scaffolds on the chromosomes. We first grouped scaffolds together on the basis of their order in the components from CSA. These grouped scaffolds were reordered by examining residual mate-pairing data between the scaffolds. We next mapped the scaffold groups onto the chromosome using physical mapping data. This step depends on having reliable high-resolution map information such that each scaffold will overlap multiple markers. There are two genome-wide types of map information available: high-density STS maps and fingerprint maps of BAC clones developed at Washington University (*45*). Among the genome-wide STS maps, GeneMap99 (GM99) has the most markers and therefore was most useful for mapping scaffolds. The two different mapping approaches are complementary to one another. The fingerprint maps should have better local order because they were built by comparison of overlapping BAC clones. On the other hand, GM99 should have a more reliable long-range order, because the framework markers were derived from well-validated genetic maps. Both types of maps were used as a reference for human curation of the components that were the input to the regional assembly, but they did not determine the order of sequences produced by the assembler.

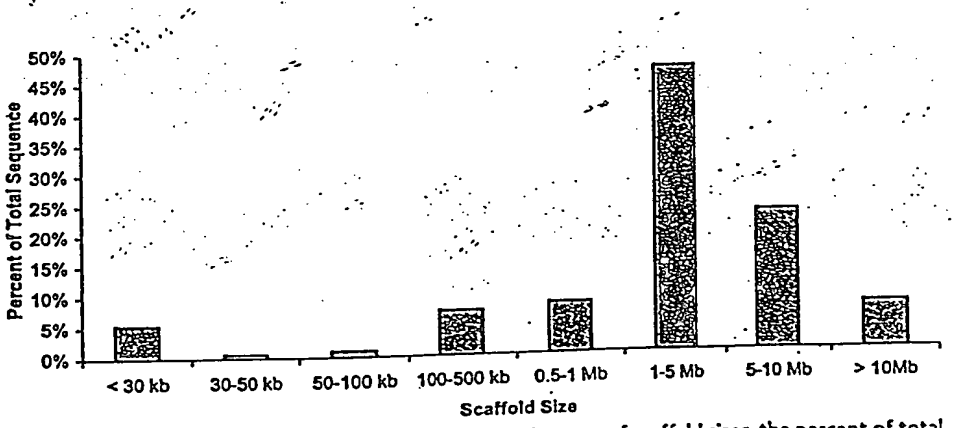


Fig. 5. Distribution of scaffold sizes of the CSA. For each range of scaffold sizes, the percent of total sequence is indicated.

In order to determine the effectiveness of the fingerprint maps and GM99 for mapping scaffolds, we first examined the reliability of these maps by comparison with large scaffolds. Only 1% of the STS markers on the 10 largest scaffolds (those >9 Mbp) were mapped on a different chromosome on GM99. Two percent of the STS markers disagreed in position by more than five framework bins. However, for the fingerprint maps, a 2% chromosome discrepancy was observed, and on average 23.8% of BAC locations in the scaffold sequence disagreed with fingerprint map placement by more than five BACs. When further examining the source of discrepancy, it was found that most of the discrepancy came from 4 of the 10 scaffolds, indicating this there is variation in the quality of either the map or the scaffolds. All four scaffolds were assembled, as well as the other six, as judged by clone coverage analysis, and showed the same low discrepancy rate to GM99, and thus we concluded that the fingerprint map global order in these cases was not reliable. Smaller scaffolds had a higher discordance rate with GM99 (4.21% of STSs were discordant by more than five framework bins), but a lower discordance rate with the fingerprint maps (11% of BACs disagreed with fingerprint maps by more than five BACs). This observation agrees with the clone coverage analysis (*46*) that Celera scaffold construction was better supported by long-range mate pairs in larger scaffolds than in small scaffolds.

We created two orderings of Celera scaffolds on the basis of the markers (BAC or STS) on these maps. Where the order of scaffolds agreed between GM99 and the WashU BAC map, we had a high degree of confidence that that order was correct; these scaffolds were termed "anchor scaffolds." Only scaffolds with a low overall discrepancy rate with both maps were considered anchor scaffolds. Scaffolds in GM99 bins were allowed to permute in their order to match WashU ordering, provided they did not violate their framework orders. Orientation of individual scaffolds was determined by the presence of multiple mapped markers with consistent order. Scaffolds with only one marker have insufficient information to assign orientation. We found 70.1% of the genome in anchored scaffolds, more than 99% of which are also oriented (Table 4). Because GM99 is of lower resolution than the WashU map, a number of scaffolds without STS matches could be ordered relative to the anchored scaffolds because they included sequence from the same or adjacent BACs on the WashU map. On the other hand, because of occasional WashU global ordering discrepancies, a number of scaffolds determined to be "unmappable" on the WashU map could be ordered relative to the anchored scaffolds

with GM99. These scaffolds were termed "ordered scaffolds." We found that 13.9% of the assembly could be ordered by these additional methods, and thus 84.0% of the genome was ordered unambiguously.

Next, all scaffolds that could be placed, but not ordered, between anchors were assigned to the interval between the anchored scaffolds and were deemed to be "bounded" between them. For example, small scaffolds having STS hits from the same GeneMap bin or hitting the same BAC cannot be ordered relative to each other, but can be assigned a placement boundary relative to other anchored or ordered scaffolds. The remaining scaffolds either had no localization information, conflicting information, or could only be assigned to a generic chromosome location. Using the above approaches, ~98% of the genome was anchored, ordered, or bounded.

Finally, we assigned a location for each scaffold placed on the chromosome by spreading out the scaffolds per chromosome. We assumed that the remaining unmapped scaffolds, constituting 2% of the genome, were distributed evenly across the genome. By dividing the sum of unmapped scaffold lengths with the sum of the number of mapped scaffolds, we arrived at an estimate of interscaffold gap of 1483 bp. This gap was used to separate all the scaffolds on each chromosome and to assign an offset in the chromosome.

During the scaffold-mapping effort, we encountered many problems that resulted in additional quality assessment and validation analysis. At least 978 (3% of 33,173) BACs were believed to have sequence data from more than one location in the genome (*47*). This is consistent with the bactig chimerism analysis reported above in the Assembly Strategies section. These BACs could not be assigned to unique positions within the CSA assembly and thus could not be used for ordering scaffolds. Likewise, it was not always possible to assign STSs to unique locations in the assembly because of genome duplications, repetitive elements, and pseudogenes.

Because of the time required for an exhaustive search for a perfect overlap, CSA generated 21,607 intrascaffold gaps where the mate-pair data suggested that the contigs should overlap, but no overlap was found. These gaps were defined as a fixed 50 bp in length and make up 18.6% of the total 116,442 gaps in the CSA assembly.

We chose not to use the order of exons implied in cDNA or EST data as a way of ordering scaffolds. The rationale for not using this data was that doing so would have biased certain regions of the assembly by rearranging scaffolds to fit the transcript data and made validation of both the assembly and gene definition processes more difficult.

4.7 Assembly and validation analysis

We analyzed the assembly of the genome from the perspectives of completeness (amount of coverage of the genome) and correctness (the structural accuracy of the order and orientation and the consensus sequence of the assembly).

Completeness. Completeness is defined as the percentage of the euchromatic sequence represented in the assembly. This cannot be known with absolute certainty until the euchromatin sequence has been completed. However, it is possible to estimate completeness on the basis of (i) the estimated sizes of intrascaffold gaps; (ii) coverage of the two published chromosomes, 21 and 22 (*48*, *49*); and (iii) analysis of the percentage of an independent set of random sequences (STS markers) contained in the assembly. The whole-genome libraries contain heterochromatic sequence and, although no attempt has been made to assemble it, there may be instances of unique sequence embedded in regions of heterochromatin as were observed in *Drosophila* (*50*, *51*).

The sequences of human chromosomes 21 and 22 have been completed to high quality and published (*48*, *49*). Although this sequence served as input to the assembler, the finished sequence was shredded into a shotgun data set so that the assembler had the opportunity to assemble it differently from the original sequence in the case of structural polymorphisms or assembly errors in the BAC data. In particular, the assembler must be able to resolve repetitive elements at the scale of components (generally multimegabase in size), and so this comparison reveals the level to which the assembler resolves repeats. In certain areas, the assembly structure differs from the published versions of chromosomes 21 and 22 (see below). The consequence of the flexibility to assemble "finished" sequence differently on the basis of Celera data resulted in an assembly with more segments than the chromosome 21 and 22 sequences. We examined the reasons why there are more gaps in the Celera sequence than in chromosomes 21 and 22 and expect that they may be typical of gaps in other regions of the genome. In the Celera assembly, there are 25 scaffolds, each containing at least 10 kb of sequence, that collectively span 94.3% of chromosome 21. Sixty-two scaffolds span 95.7% of chromosome 22. The total length of the gaps remaining in the Celera assembly for these two chromosomes is 3.4 Mbp. These gap sequences were analyzed by RepeatMasker and by searching against the entire genome assembly (*52*). About 50% of the gap sequence consisted of common repetitive elements identified by RepeatMasker; more than half of the remainder was lower copy number repeat elements.

A more global way of assessing completeness is to measure the content of an independent set of sequence data in the assembly. We compared 48,938 STS markers from Genemap99 (*51*) to the scaffolds. Because these markers were not used in the assembly processes, they provided a truly independent measure of completeness. ePCR (*53*) and BLAST (*54*) were used to locate STSs on the assembled genome. We found 44,524 (91%) of the STSs in the mapped genome. An additional 2648 markers (5.4%) were found by searching the unassembled data or "chaff." We identified 1283 STS markers (2.6%) not found in either Celera sequence or BAC data as of September 2000, raising the possibility that these markers may not be of human origin. If that were the case, the Celera assembled sequence would represent 93.4% of the human genome and the unassembled data 5.5%, for a total of 98.9% coverage. Similarly, we compared CSA against 36,678 TNG radiation hybrid markers (*55a*) using the same method. We found that 32,371 markers (88%) were located in the mapped CSA scaffolds, with 2055 markers (5.6%) found in the remainder. This gave a 94% coverage of the genome through another genome-wide survey.

Correctness. Correctness is defined as the structural and sequence accuracy of the assembly. Because the source sequences for the Celera data and the GenBank data are from different individuals, we could not directly compare the consensus sequence of the as-

Table 4. Summary of scaffold mapping. Scaffolds were mapped to the genome with different levels of confidence (anchored scaffolds have the highest confidence; unmapped scaffolds have the lowest). Anchored scaffolds were consistently ordered by the WashU BAC map and GM99. Ordered scaffolds were consistently ordered by at least one of the following: the WashU BAC map, GM99, or component tiling path. Bounded scaffolds had order conflicts between at least two of the external maps, but their placements were adjacent to a neighboring anchored or ordered scaffold. Unmapped scaffolds had, at most, a chromosome assignment. The scaffold subcategories are given below each category.

Mapped scaffold category	Number	Length (bp)	% Total length
Anchored	1,526	1,860,676,676	70
Oriented	1,246	1,852,088,645	70
Unoriented	280	8,588,031	0.3
Ordered	2,001	369,235,857	14
Oriented	839	329,633,166	12
Unoriented	1,162	39,602,691	2
Bounded	38,241	368,753,463	14
Oriented	7,453	274,536,424	10
Unoriented	30,788	94,217,039	4
Unmapped	11,823	55,313,737	2
Known chromosome	281	2,505,844	0.1
Unknown chromosome	11,542	52,807,893	2

sembly against other finished sequence for determining sequencing accuracy at the nucleotide level, although this has been done for identifying polymorphisms as described in Section 6. The accuracy of the consensus sequence is at least 99.96% on the basis of a statistical estimate derived from the quality values of the underlying reads.

The structural consistency of the assembly can be measured by mate-pair analysis. In a correct assembly, every mated pair of sequencing reads should be located on the consensus sequence with the correct separation and orientation between the pairs. A pair is termed "valid" when the reads are in the correct orientation and the distance between them is within the mean ± 3 standard deviations of the distribution of insert sizes of the library from which the pair was sampled. A pair is termed "misoriented" when the reads are not correctly oriented, and is termed "misseparated" when the distance between the reads is not in the correct range but the reads are correctly oriented. The mean ± the standard deviation of each library used by the assembler was determined as described above. To validate these, we examined all reads mapped to the finished sequence of chromosome 21 (*48*) and determined how many incorrect mate pairs there were as a result of laboratory tracking errors and chimerism (two different segments of the genome cloned into the same plasmid), and how tight the distribution of insert sizes was for those that were correct (Table 5). The standard deviations for all Celera libraries were quite small, less than 15% of the insert length, with the exception of a few 50-kbp libraries. The 2- and 10-kbp libraries contained less than 2% invalid mate pairs, whereas the 50-kbp libraries were somewhat higher (~10%). Thus, although the mate-pair information was not perfect, its accuracy was such that measuring valid, misoriented, and misseparated pairs with respect to a given assembly was deemed to be a reliable instrument for validation purposes, especially when several mate pairs confirm or deny an ordering.

The clone coverage of the genome was 39×, meaning that any given base pair was, on average, contained in 39 clones or, equivalently, spanned by 39 mate-paired reads. Areas of low clone coverage or areas with a high proportion of invalid mate pairs would indicate potential assembly problems. We computed the coverage of each base in the assembly by valid mate pairs (Table 6). In summary, for scaffolds >30 kbp in length, less than 1% of the Celera assembly was in regions of less than 3× clone coverage. Thus, more than 99% of the assembly, including order and orientation, is strongly supported by this measure alone.

We examined the locations and number of all misoriented and misseparated mates. In addition to doing this analysis on the CSA assembly (as of 1 October 2000), we also performed a study of the PFP assembly as of 5 September 2000 (*30*, *55b*). In this latter case, Celera mate pairs had to be mapped to the PFP assembly. To avoid mapping errors due to high-fidelity repeats, the only pairs mapped were those for which both reads matched at only one location with less than 6% differences. A threshold was set such that sets of five or more simultaneously invalid mate pairs indicated a potential breakpoint, where the construction of the two assemblies differed. The graphic comparison of the CSA chromosome 21 assembly with the published sequence (Fig. 6A) serves as a validation of this methodology. Blue tick marks in the panels indicate breakpoints. There were a similar (small) number of breakpoints on both chromosome sequences. The exception was 12 sets of scaffolds in the Celera assembly (a total of 3% of the chromosome length in 212 single-contig scaffolds) that were mapped to the wrong positions because they were too small to be mapped reliably. Figures 6 and 7 and Table 6 illustrate the mate-pair differences and breakpoints between the two assemblies. There was a higher percentage of misoriented and misseparated mate pairs in the large-insert libraries (50 kbp and BAC ends) than in the small-insert libraries in both assemblies (Table 6). The large-insert libraries are more likely to identify discrepancies simply because they span a larger segment of the genome. The graphic comparison between the two assemblies for chromosome 8 (Fig. 6, B and C) shows that there are many

Table 5. Mate-pair validation. Celera fragment sequences were mapped to the published sequence of chromosome 21. Each mate pair uniquely mapped was evaluated for correct orientation and placement (number of mate pairs tested). If the two mates had incorrect relative orientation or placement, they were considered invalid (number of invalid mate pairs).

		Chromosome 21						Genome		
Library type	Library no.	Mean insert size (bp)	SD (bp)	SD/ mean (%)	No. of mate pairs tested	No. of invalid mate pairs	% invalid	Mean insert size (bp)	SD (bp)	SD/ mean (%)
2 kbp	1	2,081	106	5.1	3,642	38	1.0	2,082	90	4.3
	2	1,913	152	7.9	28,029	413	1.5	1,923	118	6.1
	3	2,166	175	8.1	4,405	57	1.3	2,162	158	7.3
10 kbp	4	11,385	851	7.5	4,319	80	1.9	11,370	696	6.1
	5	14,523	1,875	12.9	7,355	156	2.1	14,142	1,402	9.9
	6	9,635	1,035	10.7	5,573	109	2.0	9,606	934	9.7
	7	10,223	928	9.1	34,079	399	1.2	10,190	777	7.6
50 kbp	8	64,888	2,747	4.2	16	1	6.3	65,500	5,504	8.4
	9	53,410	5,834	10.9	914	170	18.6	53,311	5,546	10.4
	10	52,034	7,312	14.1	5,871	569	9.7	51,498	6,588	12.8
	11	52,282	7,454	14.3	2,629	213	8.1	52,282	7,454	14.3
	12	46,616	7,378	15.8	2,153	215	10.0	45,418	9,068	20.0
	13	55,788	10,099	18.1	2,244	249	11.1	53,062	10,893	20.5
	14	39,894	5,019	12.6	199	7	3.5	36,838	9,988	27.1
BES	15	48,931	9,813	20.1	144	10	6.9	47,845	4,774	10.0
	16	48,130	4,232	8.8	195	14	7.2	47,924	4,581	9.6
	17	106,027	27,778	26.2	330	16	4.8	152,000	26,600	17.5
	18	160,575	54,973	34.2	155	8	5.2	161,750	27,000	16.7
	19	164,155	19,453	11.9	642	44	6.9	176,500	19,500	11.05
Sum					102,894	2,768 (mean = 2.7)	2.7			

more breakpoints for the PFP assembly than for the Celera assembly. Figure 7 shows the breakpoint map (blue tick marks) for both assemblies of each chromosome in a side-by-side fashion. The order and orientation of Celera's assembly shows substantially fewer breakpoints except on the two finished chromosomes. Figure 7 also depicts large gaps (>10 kbp) in both assemblies as red tick marks. In the CSA assembly, the size of all gaps have been estimated on the basis of the mate-pair data. Breakpoints can be caused by structural polymorphisms, because the two assemblies were derived from different human genomes. They also reflect the unfinished nature of both genome assemblies.

3 Gene Prediction and Annotation

Summary. To enumerate the gene inventory, we developed an integrated, evidence-based approach named Otto. The evidence used to increase the likelihood of identifying genes includes regions conserved between the mouse and human genomes, similarity to ESTs or other mRNA-derived data, or similarity to other proteins. A comparison of Otto (combined Otto-RefSeq and Otto homology) with Genscan, a standard gene-prediction algorithm, showed greater sensitivity (0.78 versus 0.50) and specificity (0.93 versus 0.63) of Otto in the ability to define gene structure. Otto-predicted genes were complemented with a set of genes from three gene-prediction programs that exhibited weaker, but still significant, evidence that they may be expressed. Conservative criteria, requiring at least two lines of evidence, were used to define a set of 26,383 genes with good confidence that were used for more detailed analysis presented in the subsequent sections. Extensive manual curation to establish precise characterization of gene structure will be necessary to improve the results from this initial computational approach.

3.1 Automated gene annotation

A gene is a locus of cotranscribed exons. A single gene may give rise to multiple transcripts, and thus multiple distinct proteins with multiple functions, by means of alternative splicing and alternative transcription initiation and termination sites. Our cells are able to discern within the billions of base pairs of the genomic DNA the signals for initiating transcription and for splicing together exons separated by a few or hundreds of thousands of base pairs. The first step in characterizing the genome is to define the structure of each gene and each transcription unit.

The number of protein-coding genes in mammals has been controversial from the outset. Initial estimates based on reassociation data placed it between 30,000 to 40,000, whereas later estimates from the brain were >100,000 (*56*). More recent data from both the corporate and public sectors, based on extrapolations from EST, CpG island, and transcript density–based extrapolations, have not reduced this variance. The highest recent number of 142,634 genes emanates from a report from Incyte Pharmaceuticals, and is based on a combination of EST data and the association of ESTs with CpG islands (*57*). In stark contrast are three quite different, and much lower estimates: one of ~35,000 genes derived with genome-wide EST data and sampling procedures in conjunction with chromosome 22 data (*58*); another of 28,000 to 34,000 genes derived with a comparative methodology involving sequence conservation between humans and the puffer fish *Tetraodon nigroviridis* (*59*); and a figure of 35,000 genes, which was derived simply by extrapolating from the density of 770 known and predicted genes in the 67 Mbp of chromosomes 21 and 22, to the approximately 3-Gbp euchromatic genome.

The problem of computational identification of transcriptional units in genomic DNA sequence can be divided into two phases. The first is to partition the sequence into segments that are likely to correspond to individual genes. This is not trivial and is a weakness of most de novo gene-finding algorithms. It is also critical to determining the number of genes in the human gene inventory. The second challenge is to construct a gene model that reflects the probable structure of the transcript(s) encoded in the region. This can be done with reasonable accuracy when a full-length cDNA has been sequenced or a highly homologous protein sequence is known. De novo gene prediction, although less accurate, is the only way to find genes that are not represented by homologous proteins or ESTs. The following section describes the methods we have developed to address these problems for the prediction of protein-coding genes.

We have developed a rule-based expert system, called Otto, to identify and characterize genes in the human genome (*60*). Otto attempts to simulate in software the process that a human annotator uses to identify a gene and refine its structure. In the process of annotating a region of the genome, a human curator examines the evidence provided by the computational pipeline (described below) and examines how various types of evidence relate to one another. A curator puts different levels of confidence in different types of evidence and looks for certain patterns of evidence to support gene annotation. For example, a curator may examine homology to a number of ESTs and evaluate whether or not they can be connected into a longer, virtual mRNA. The curator would also evaluate the strength of the similarity and the contiguity of the match, in essence asking whether any ESTs cross splice-junctions and whether the edges of putative exons have consensus splice sites. This kind of manual annotation process was used to annotate the *Drosophila* genome.

The Otto system can promote observed evidence to a gene annotation in one of two ways. First, if the evidence includes a high-quality match to the sequence of a known gene [here defined as a human gene represented in a curated subset of the RefSeq database (*61*)], then Otto can promote this to a gene annotation. In the second method, Otto evaluates a broad spectrum of evidence and determines if this evidence is adequate to support promotion to a gene annotation. These processes are described below.

Initially, gene boundaries are predicted on the basis of examination of sets of overlapping protein and EST matches generated by a computational pipeline (*62*). This pipeline searches the scaffold sequences against protein, EST, and genome-sequence databases to define regions of sequence similarity and runs three de novo gene-prediction programs.

To identify likely gene boundaries, regions of the genome were partitioned by Otto on the basis of sequence matches identified by BLAST. Each of the database sequences matched in the region under analysis was compared by an algorithm that takes into account both coordinates of the matching sequence, as well as the sequence type (e.g., protein, EST, and so forth). The results were used to group the matches into bins of related sequences that may define a gene and identi-

Table 6. Genome-wide mate pair analysis of compartmentalized shotgun (CSA) and PFP assemblies.*

Genome library	CSA			PFP		
	% valid	% mis-oriented	% mis-separated†	% valid	% mis-oriented	% mis-separated†
2 kbp	98.5	0.6	1.0	95.7	2.0	2.3
10 kbp	96.7	1.0	2.3	81.9	9.6	8.6
50 kbp	93.9	4.5	1.5	64.2	22.3	13.5
BES	94.1	2.1	3.8	62.0	19.3	18.8
Mean	97.4	1.0	1.6	87.3	6.8	5.9

*Data for individual chromosomes can be found in Web fig. 3 on *Science* Online at www.sciencemag.org/cgi/content/full/291/5507/1304/DC1. †Mates are misseparated if their distance is >3 SD from the mean library size.

gene boundaries. During this process, multiple hits to the same region were collapsed to a coherent set of data by tracking the coverage of a region. For example, if a group of bases was represented by multiple overlapping ESTs, the union of these regions matched by the set of ESTs on the scaffold was marked as being supported by EST evidence. This resulted in a series of "gene bins," each of which was believed to contain a single gene. One weakness of this initial implementation of the algorithm was in predicting gene boundaries in regions of tandemly duplicated genes. Gene clusters frequently resulted in homologous neighboring genes being joined together, resulting in an annotation that artificially concatenated these gene models.

Next, known genes (those with exact matches of a full-length cDNA sequence to the genome) were identified, and the region corresponding to the cDNA was annotated as a predicted transcript. A subset of the curated human gene set RefSeq from the National Center for Biotechnology Information (NCBI) was included as a data set searched in the computational pipeline. If a RefSeq transcript matched the genome assembly for at least 50% of its length at >92% identity, then the SIM4 (*63*) alignment of the RefSeq transcript to the region of the genome under analysis was promoted to the status of an Otto annotation. Because the genome sequence has gaps and sequence errors such as frameshifts, it was not always possible to predict a transcript that agrees precisely with the experimentally determined cDNA sequence. A total of 6538 genes in our inventory were identified and transcripts predicted in this way.

Regions that have a substantial amount of sequence similarity, but do not match known genes, were analyzed by that part of the Otto system that uses the sequence similarity information to predict a transcript. Here, Otto

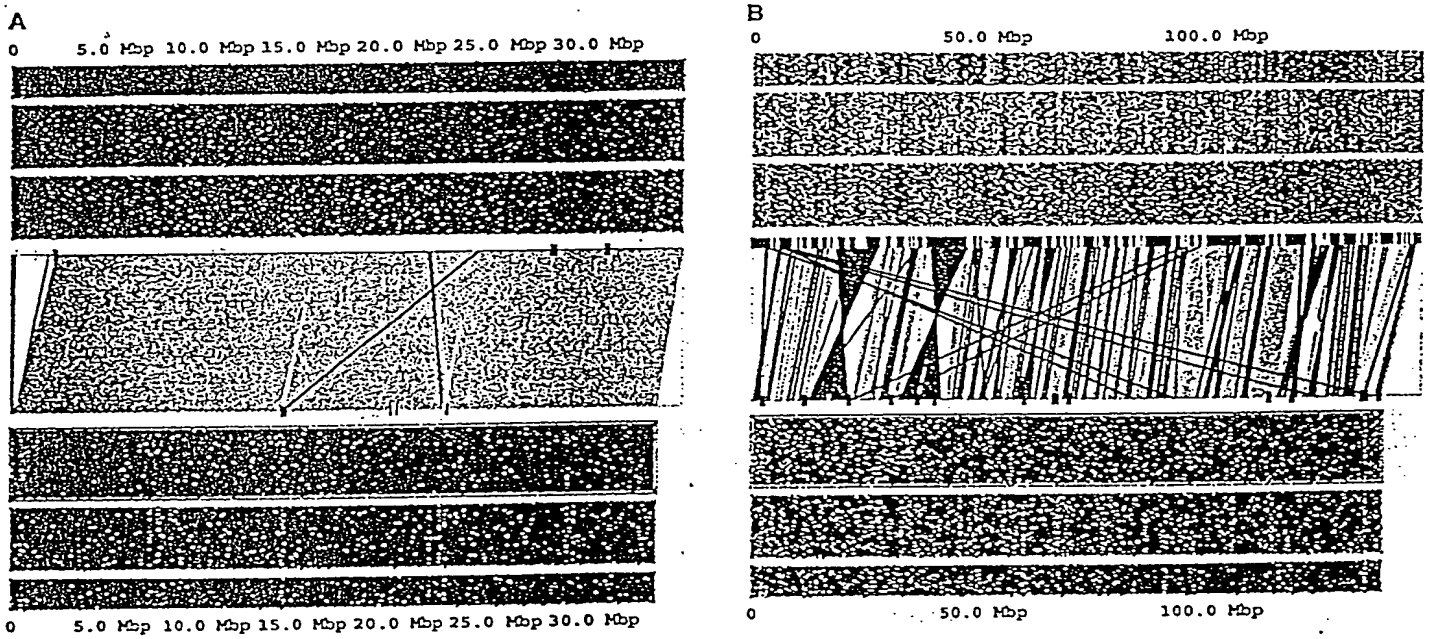


Fig. 6. Comparison of the CSA and the PFP assembly. (A) All of chromosome 21, (B) all of chromosome 8, and (C) a 1-Mb region of chromosome 8 representing a single Celera scaffold. To generate the figure, Celera fragment sequences were mapped onto each assembly. The PFP assembly is indicated in the upper third of each panel; the Celera assembly is indicated in the lower third. In the center of the panel, green lines show Celera sequences that are in the same order and orientation in both assemblies and form the longest consistently ordered run of sequences. Yellow lines indicate sequence blocks that are in the same orientation, but out of order. Red lines indicate sequence blocks that are not in the same orientation. For clarity, in the latter two cases, lines are only drawn between segments of matching sequence that are at least 50 kbp long. The top and bottom thirds of each panel show the extent of Celera mate-pair violations (red, misoriented; yellow, incorrect distance between the mates) for each assembly grouped by library size. (Mate pairs that are within the correct distance, as expected from the mean library insert size, are omitted from the figure for clarity.) Predicted breakpoints, corresponding to stacks of violated mate pairs of the same type, are shown as blue ticks on each assembly axis. Runs of more than 10,000 Ns are shown as cyan bars. Plots of all 24 chromosomes can be seen in Web fig. 3 on *Science* Online at www.sciencemag.org/cgi/content/full/291/5507/1304/DC1.

evaluates evidence generated by the computational pipeline, corresponding to conservation between mouse and human genomic DNA, similarity to human transcripts (ESTs and cDNAs), similarity to rodent transcripts (ESTs and cDNAs), and similarity of the translation of human genomic DNA to known proteins to predict potential genes in the human genome. The sequence from the region of genomic DNA contained in a gene bin was extracted, and the subsequences supported by any homology evidence were marked (plus 100

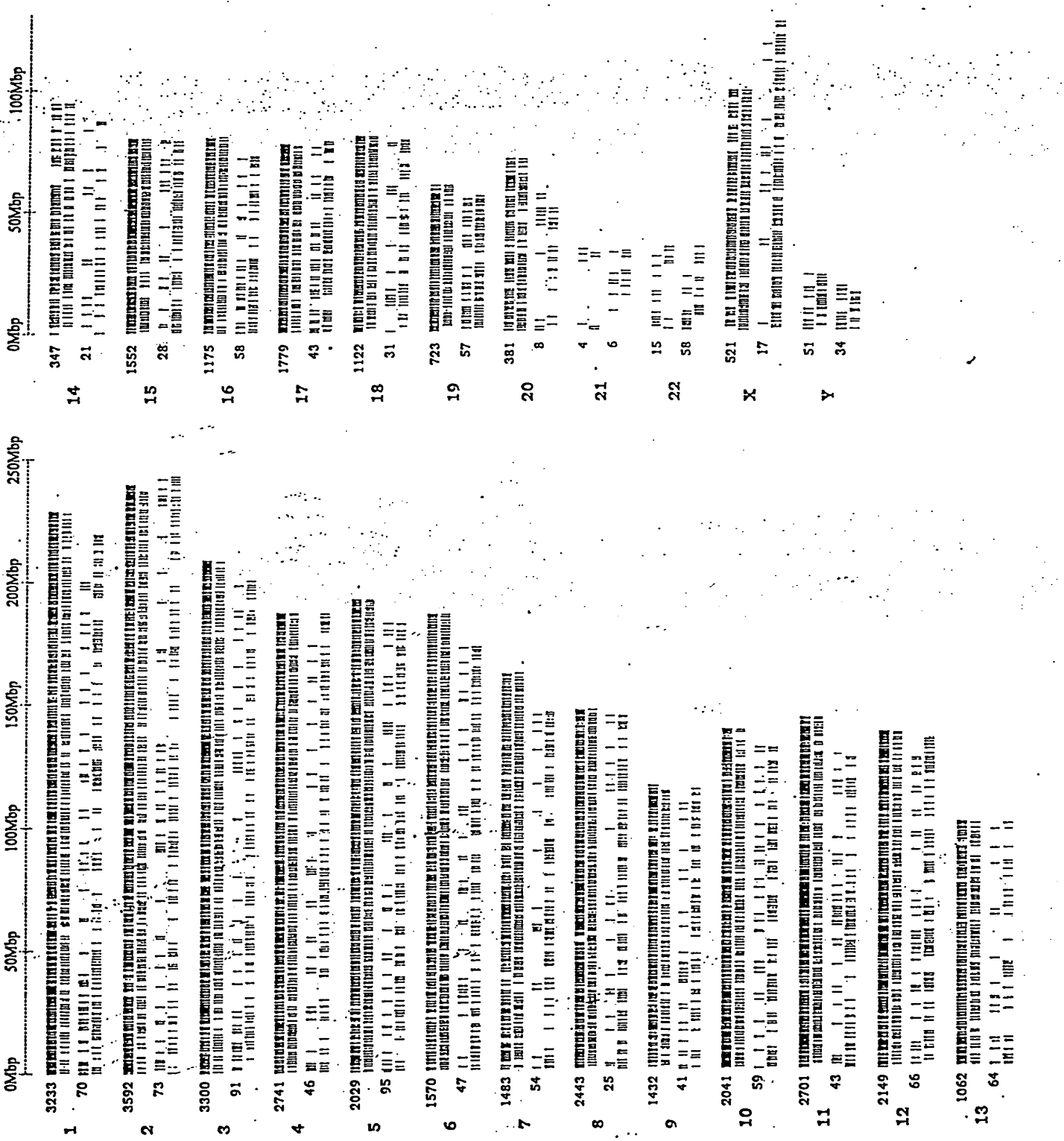


Fig. 7. Schematic view of the distribution of breakpoints and large gaps on all chromosomes. For each chromosome, the upper pair of lines represent the PFP assembly, and the lower pair of lines represent Celera's assembly. Blue tick marks represent breakpoints, whereas red tick marks represent a gap of larger than 10,000 bp. The number of breakpoints per chromosome is indicated in black, and the chromosome numbers in red.

bases flanking these regions). The other bases in the region, those not covered by any homology evidence, were replaced by N's. This sequence segment, with high confidence regions represented by the consensus genomic sequence and the remainder represented by N's, was then evaluated by Genscan to see if a consistent gene model could be generated. This procedure simplified the gene-prediction task by first establishing the boundary for the gene (not a strength of most gene-finding algorithms), and by eliminating regions with no supporting evidence. If Genscan returned a plausible gene model, it was further evaluated before being promoted to an "Otto" annotation. The final Genscan predictions were often quite different from the prediction that Genscan returned on the same region of native genomic sequence. A weakness of using Genscan to refine the gene model is the loss of valid, small exons from the final annotation.

The next step in defining gene structures based on sequence similarity was to compare each predicted transcript with the homology-based evidence that was used in previous steps to evaluate the depth of evidence for each exon in the prediction. Internal exons were considered to be supported if they were covered by homology evidence to within ±10 bases of their edges. For first and last exons, the internal edge was required to be within 10 bases, but the external edge was allowed greater latitude to allow for 5′ and 3′ untranslated regions (UTRs). To be retained, a prediction for a multi-exon gene must have evidence such that the total number of "hits," as defined above, divided by the number of exons in the prediction must be >0.66 or must correspond to a RefSeq sequence. A single-exon gene must be covered by at least three supporting hits (±10 bases on each side), and these must cover the complete predicted open reading frame. For a single-exon gene, we also required that the Genscan prediction include both a start and a stop codon. Gene models that did not meet these criteria were disregarded, and those that passed were promoted to Otto predictions. Homology-based Otto predictions do not contain 3′ and 5′ untranslated sequence. Although three de novo gene-finding programs [GRAIL, Genscan, and FgenesH (*63*)] were run as part of the computational analysis, the results of these programs were not directly used in making the Otto predictions. Otto predicted 11,226 additional genes by means of sequence similarity.

Table 7. Sensitivity and specificity of Otto and Genscan. Sensitivity and specificity were calculated by first aligning the prediction to the published RefSeq transcript, tallying the number (*N*) of uniquely aligned RefSeq bases. Sensitivity is the ratio of *N* to the length of the published RefSeq transcript. Specificity is the ratio of *N* to the length of the prediction. All differences are significant (Tukey HSD; $P < 0.001$).

Method	Sensitivity	Specificity
Otto (RefSeq only)*	0.939	0.973
Otto (homology)†	0.604	0.884
Genscan	0.501	0.633

*Refers to those annotations produced by Otto using only the Sim4-polished RefSeq alignment rather than an evidence-based Genscan prediction. †Refers to those annotations produced by supplying all available evidence to Genscan.

3.2 Otto validation

To validate the Otto homology-based process and the method that Otto uses to define the structures of known genes, we compared transcripts predicted by Otto with their corresponding (and presumably correct) transcript from a set of 4512 RefSeq transcripts for which there was a unique SIM4 alignment (Table 7). In order to evaluate the relative performance of Otto and Genscan, we made three comparisons. The first involved a determination of the accuracy of gene models predicted by Otto with only homology data other than the corresponding RefSeq sequence (Otto homology in Table 7). We measured the sensitivity (correctly predicted bases divided by the total length of the cDNA) and specificity (correctly predicted bases divided by the sum of the correctly and incorrectly predicted bases). Second, we examined the sensitivity and specificity of the Otto predictions that were made solely with the RefSeq sequence, which is the process that Otto uses to annotate known genes (Otto-RefSeq). And third, we determined the accuracy of the Genscan predictions corresponding to these RefSeq sequences. As expected, the alignment method (Otto-RefSeq) was the most accurate, and Otto-homology performed better than Genscan by both criteria. Thus, 6.1% of true RefSeq nucleotides were not represented in the Otto-refseq annotations and 2.7% of the nucleotides in the Otto-RefSeq transcripts were not contained in the original RefSeq transcripts. The discrepancies could come from legitimate differences between the Celera assembly and the RefSeq transcript due to polymorphisms, incomplete or incorrect data in the Celera assembly, errors introduced by Sim4 during the alignment process, or the presence of alternatively spliced forms in the data set used for the comparisons.

Because Otto uses an evidence-based approach to reconstruct genes, the absence of experimental evidence for intervening exons may inadvertantly result in a set of exons that cannot be spliced together to give rise to a transcript. In such cases, Otto may "split genes" when in fact all the evidence should be combined into a single transcript. We also examined the tendency of these methods to incorrectly split gene predictions. These trends are shown in Fig. 8. Both RefSeq and homology-based predictions by Otto split known genes into fewer segments than Genscan alone.

3.3 Gene number

Recognizing that the Otto system is quite conservative, we used a different gene-prediction strategy in regions where the homology evidence was less strong. Here the results of de novo gene predictions were used. For these genes, we insisted that a predicted transcript have at least two of the following types of evidence to be included in the gene set for further analysis: protein, human EST, rodent EST, or mouse genome fragment matches. This final class of predicted genes is a subset of the predictions made by the three gene-finding programs that were used in the computational pipeline. For these, there was not sufficient sequence similarity information for Otto to attempt to predict a gene structure. The three de novo gene-finding programs resulted in about 155,695 predictions, of which ~76,410 were nonredundant (nonoverlapping with one another). Of these, 57,935 did not overlap known genes or predictions made by Otto. Only 21,350 of the gene predictions that did not overlap Otto predictions were partially supported by at least one type of sequence similarity evidence, and 8619 were partially supported by two types of evidence (Table 8).

The sum of this number (21,350) and the number of Otto annotations (17,764), 39,114, is near the upper limit for the human gene complement. As seen in Table 8, if the requirement for other supporting evidence is made more stringent, this number drops rapidly so that demanding two types of evidence reduces the total gene number to 26,383 and demanding three types reduces it to ~23,000. Requiring that a prediction be supported by all four categories of evidence is too stringent because it would eliminate genes that encode novel proteins (members of currently undescribed protein families). No correction for pseudogenes has been made at this point in the analysis.

In a further attempt to identify genes that were not found by the autoannotation process or any of the de novo gene finders, we examined regions outside of gene predictions that were similar to the EST sequence, and where the EST matched the genomic sequence across a splice junction. After correcting for potential 3′ UTRs of predicted genes, about 2500 such regions remained. Addition of a requirement for at least one of the following evidence types—homology to mouse genomic sequence fragments, rodent ESTs, or cDNAs—or similarity to a known protein reduced this number to 1010. Adding this to the numbers from the previous paragraph would give us estimates of about 40,000, 27,000, and 24,000 potential genes in the human genome, depending on the stringency of evidence considered. Table 8 illustrates the number of genes and presents the degree of

confidence based on the supporting evidence. Transcripts encoded by a set of 26,383 genes were assembled for further analysis. This set includes the 6538 genes predicted by Otto on the basis of matches to known genes, 11,226 transcripts predicted by Otto based on homology evidence, and 8619 from the subset of transcripts from de novo gene-prediction programs that have two types of supporting evidence. The 26,383 genes are illustrated along chromosome diagrams in Fig. 1. These are a very preliminary set of annotations and are subject to all the limitations of an automated process. Considerable refinement is still necessary to improve the accuracy of these transcript predictions. All the predictions and descriptions of genes and the associated evidence that we present are the product of completely computational processes, not expert curation. We have attempted to enumerate the genes in the human genome in such a way that we have different levels of confidence based on the amount of supporting evidence: known genes, genes with good protein or EST homology evidence, and de novo gene predictions confirmed by modest homology evidence.

3.4 Features of human gene transcripts

We estimate the average span for a "typical" gene in the human DNA sequence to be about 27,894 bases. This is based on the average span covered by RefSeq transcripts, used because it represents our highest confidence set.

The set of transcripts promoted to gene annotations varies in a number of ways. As can be seen from Table 8 and Fig. 9, transcripts predicted by Otto tend to be longer, having on average about 7.8 exons, whereas those promoted from gene-prediction programs average about 3.7 exons. The largest number of exons that we have identified in a transcript is 234 in the titin mRNA. Table 8 compares the amounts of evidence that support the Otto and other predicted transcripts. For example, one can see that a typical Otto transcript has 6.99 of its 7.81 exons supported by protein homology evidence. As would be expected, the Otto transcripts generally have more support than do transcripts predicted by the de novo methods.

4 Genome Structure

Summary. This section describes several of the noncoding attributes of the assembled genome sequence and their correlations with the predicted gene set. These include an analysis of G+C content and gene density in the context of cytogenetic maps of the genome, an enumerative analysis of CpG islands, and a brief description of the genome-wide repetitive elements.

4.1 Cytogenetic maps

Perhaps the most obvious, and certainly the most visible, element of the structure of the genome is the banding pattern produced by Giemsa stain. Chromosomal banding studies have revealed that about 17% to 20% of the human chromosome complement consists of C-bands, or constitutive heterochromatin (*64*). Much of this heterochromatin is highly polymorphic and consists of different families of alpha satellite DNAs with various higher order repeat structures (*65*). Many chromosomes have complex inter- and intrachromosomal duplications present in pericentromeric regions (*66*). About 5% of the sequence reads were identified as alpha satellite sequences; these were not included in the assembly.

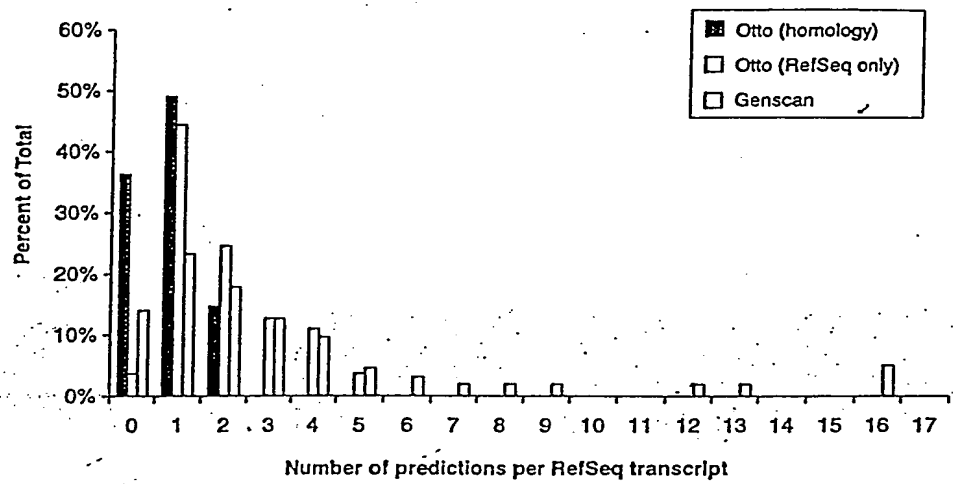


Fig. 8. Analysis of split genes resulting from different annotation methods. A set of 4512 Sim4-based alignments of RefSeq transcripts to the genomic assembly were chosen (see the text for criteria), and the numbers of overlapping Genscan, Otto (RefSeq only) annotations based solely on Sim4-polished RefSeq alignments, and Otto (homology) annotations (annotations produced by supplying all available evidence to Genscan) were tallied. These data show the degree to which multiple Genscan predictions and/or Otto annotations were associated with a single RefSeq transcript. The zero class for the Otto-homology predictions shown here indicates that the Otto-homology calls were made without recourse to the RefSeq transcript, and thus no Otto call was made because of insufficient evidence.

Table 8. Numbers of exons and transcripts supported by various types of evidence for Otto and de novo gene prediction methods. Highlighted cells indicate the gene sets analyzed in this paper (boldface, set of genes selected for protein analysis; italic, total set of accepted de novo predictions).

		Total	Types of evidence				No. of lines of evidence*			
			Mouse	Rodent	Protein	Human	≥1	≥2	≥3	≥4
Otto	Number of transcripts	17,969	17,065	14,881	15,477	16,374	17,968†	17,501	15,877	12,451
	Number of exons	141,218	111,174	89,569	108,431	118,869	140,710	127,955	99,574	59,804
De novo	Number of transcripts	58,032	14,463	5,094	8,043	9,220	*21,350*	8,619	4,947	1,904
	Number of exons	319,935	48,594	19,344	26,264	40,104	79,148	31,130	17,508	6,520
No. of exons per transcript	Otto	7.84	5.77	6.01	6.99	7.24	7.81	7.19	6.00	4.28
	De novo	5.53	3.17	3.80	3.27	4.36	3.7	3.56	3.42	3.16

*Four kinds of evidence (conservation in 3× mouse genomic DNA, similarity to human EST or cDNA, similarity to rodent EST or cDNA, and similarity to known proteins) were considered to support gene predictions from the different methods. The use of evidence is quite liberal, requiring only a partial match to a single exon of predicted transcript. †This number includes alternative splice forms of the 17,764 genes mentioned elsewhere in the text.

Examination of pericentromeric regions is ongoing.

The remaining ~80% of the genome, the euchromatic component, is divisible into G-, R-, and T-bands (*67*). These cytogenetic bands have been presumed to differ in their nucleotide composition and gene density, although we have been unable to determine precise band boundaries at the molecular level. T-bands are the most G+C- and gene-rich, and G-bands are G+C-poor (*68*). Bernardi has also offered a description of the euchromatin at the molecular level as long stretches of DNA of differing base composition, termed isochores (denoted L, H1, H2, and H3), which are >300 kbp in length (*69*). Bernardi defined the L (light) isochores as G+C-poor (<43%), whereas the H (heavy) isochores fall into three G+C-rich classes representing 24, 8, and 5% of the genome. Gene concentration has been claimed to be very low in the L isochores and 20-fold more enriched in the H2 and H3 isochores (*70*). By examining contiguous 50-kbp windows of G+C content across the assembly, we found that regions of G+C content >48% (H3 isochores) averaged 273.9 kbp in length, those with G+C content between 43 and 48% (H1+H2 isochores) averaged 202.8 kbp in length, and the average span of regions with <43% (L isochores) was 1078.6 kbp. The correlation between G+C content and gene density was also examined in 50-kbp windows along the assembled sequence (Table 9 and Figs. 10 and 11). We found that the density of genes was greater in regions of high G+C than in regions of low G+C content, as expected. However, the correlation between G+C content and gene density was not as skewed as previously predicted (*69*). A higher proportion of genes were located in the G+C-poor regions than had been expected.

Chromosomes 17, 19, and 22, which have a disproportionate number of H3-containing bands, had the highest gene density (Table 10). Conversely, of the chromosomes that we found to have the lowest gene density, X, 4, 18, 13, and Y, also have the fewest H3 bands. Chromosome 15, which also has few H3 bands, did not have a particularly low gene density in our analysis. In addition, chromosome 8, which we found to have a low gene density, does not appear to be unusual in its H3 banding.

How valid is Ohno's postulate (*71*) that mammalian genomes consist of oases of genes in otherwise essentially empty deserts? It appears that the human genome does indeed contain deserts, or large, gene-poor regions. If we define a desert as a region >500 kbp without a gene, then we see that 605 Mbp, or about 20% of the genome, is in deserts. These are not uniformly distributed over the various chromosomes. Gene-rich chromosomes 17, 19, and 22 have only about 12% of their collective 171 Mbp in deserts, whereas gene-poor chromosomes 4, 13, 18, and X have 27.5% of their 492 Mbp in deserts (Table 11). The apparent lack of predicted genes in these regions does not necessarily imply that they are devoid of biological function.

4.2 Linkage map

Linkage maps provide the basis for genetic analysis and are widely used in the study of the inheritance of traits and in the positional cloning of genes. The distance metric, centimorgans (cM), is based on the recombination rate between homologous chromosomes during meiosis. In general, the rate of recombination in females is greater than that in males, and this degree of map expansion is not uniform across the genome (*72*). One of the opportunities enabled by a nearly complete genome sequence is to produce the ultimate physical map, and to fully analyze its correspondence with two other maps that have been widely used in genome and genetic analysis: the linkage map and the cytogenetic map. This would close the loop between the mapping and sequencing phases of the genome project.

We mapped the location of the markers that constitute the Genethon linkage map to the genome. The rate of recombination, expressed as cM per Mbp, was calculated for 3-Mbp windows as shown in Table 12. Higher rates of recombination in the telomeric region of the chromosomes have been previously documented (*73*). From this mapping result, there is a difference of 4.99 between lowest rates and highest rates and the largest difference of 4.4 between males and females (4.99 to 0.47 on chromosome 16). This indicates that the variability in recombination rates among regions of the genome exceeds the differences in recombination rates between males and females. The human genome has recombination hotspots, where recombination rates vary fivefold or more over a space of 1 kbp, so the picture one gets of the magnitude of variability in recombination rate will depend on the size of the window

Table 9. Characteristics of G+C in Isochores.

Isochore	G+C (%)	Fraction of genome		Fraction of genes	
		Predicted*	Observed	Predicted*	Observed
H3	>48	5	9.5	37	24.8
H1/H2	43–48	25	21.2	32	26.6
L	<43	67	69.2	31	48.5

*The predictions were based on Bernardi's definitions (*70*) of the isochore structure of the human genome.

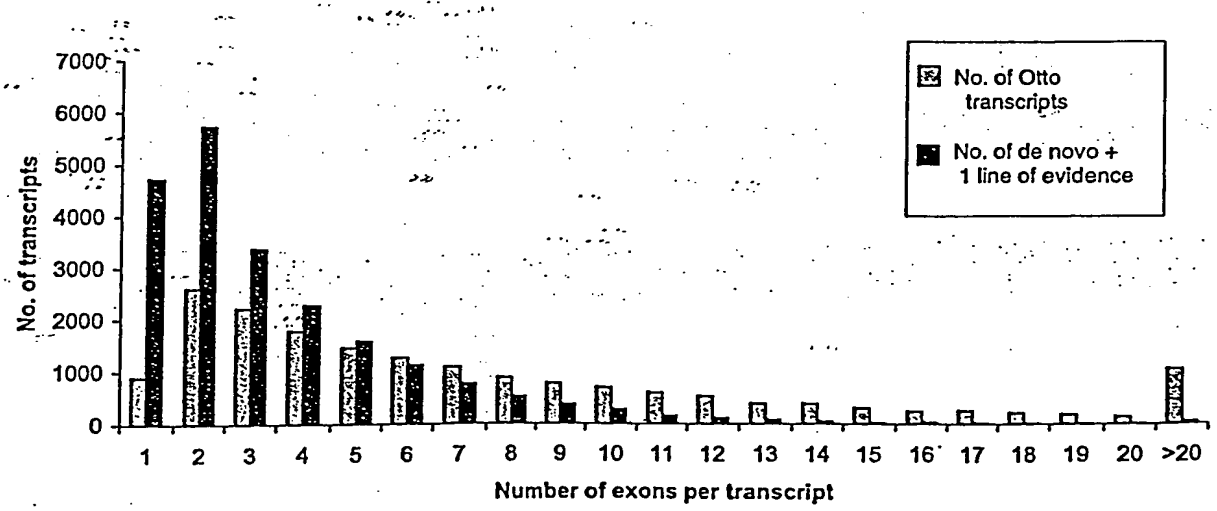


Fig. 9. Comparison of the number of exons per transcript between the 17,968 Otto transcripts and 21,350 de novo transcript predictions with at least one line of evidence that do not overlap with an Otto prediction. Both sets have the highest number of transcripts in the two-exon category, but the de novo gene predictions are skewed much more toward smaller transcripts. In the Otto set, 19.7% of the transcripts have one or two exons, and 5.7% have more than 20. In the de novo set, 49.3% of the transcripts have one or two exons, and 0.2% have more than 20.

examined. Unfortunately, too few meiotic crossovers have occurred in Centre d'Étude du Polymorphism Humain (CEPH) and other reference families to provide a resolution any finer than about 3 Mbp. The next challenge will be to determine a sequence basis of recombination at the chromosomal level. An accurate predictor for the rate for variation in recombination rates between any pair of markers would be extremely useful in designing markers to narrow a region of linkage, such as in positional cloning projects.

4.3 Correlation between CpG islands and genes

CpG islands are stretches of unmethylated DNA with a higher frequency of CpG dinucleotides when compared with the entire genome (*74*). CpG islands are believed to preferentially occur at the transcriptional start of genes, and it has been observed that most housekeeping genes have CpG islands at the 5′ end of the transcript (*75*, *76*). In addition, experimental evidence indicates that CpG island methylation is correlated with gene inactivation (*77*) and has been shown to be important during gene imprinting (*78*) and tissue-specific gene expression (*79*)

Experimental methods have been used that resulted in an estimate of 30,000 to 45,000 CpG islands in the human genome (*74*, *80*) and an estimate of 499 CpG islands on human chromosome 22 (*81*). Larsen *et al.* (*76*) and Gardiner-Garden and Frommer (*75*) used a computational method to identify CpG islands and defined them as regions of DNA of >200 bp that have a G+C content of >50% and a ratio of observed versus expected frequency of CG dinucleotide ≥0.6.

It is difficult to make a direct comparison of experimental definitions of CpG islands with computational definitions because computational methods do not consider the methylation state of cytosine and experimental methods do not directly select regions of high G+C content. However, we can determine the correlation of CpG island with gene starts, given a set of annotated genomic transcripts and the whole genome sequence. We have analyzed the publicly available annotation of chromosome 22, as well as using the entire human genome in our assembly and the computationally annotated genes. A variation of the CpG island computation was compared with Larsen *et al.* (*76*). The main differences are that we use a sliding window of 200 bp, consecutive windows are merged only if they overlap, and we recompute the CpG value upon merging, thus rejecting any potential island if it scores less than the threshold.

To compute various CpG statistics, we used two different thresholds of CG dinucleotide likelihood ratio. Besides using the original threshold of 0.6 (method 1), we used a higher threshold of CG dinucleotide likelihood ratio of 0.8 (method 2), which results in the number of CpG islands on chromosome 22 close to the number of annotated genes on this chromosome. The main results are summarized in Table 13. CpG islands computed with method 1 predicted only 2.6% of the CSA sequence as CpG, but 40% of the gene starts (start codons) are contained inside a CpG island. This is comparable to ratios reported by others (*82*). The last two rows of the table show the observed and expected average distance, respectively, of the closest CpG island from the first exon. The observed average closest CpG islands are smaller than the corresponding expected distances, confirming an association between CpG island and the first exon.

We also looked at the distribution of CpG island nucleotides among various sequence classes such as intergenic regions, introns, exons, and first exons. We computed the likelihood score for each sequence class as the ratio of the observed fraction of CpG island nucleotides in that sequence class and the expected fraction of CpG island nucleotides in that sequence class. The result of applying method 1 on CSA were scores of 0.89 for intergenic region, 1.2 for intron, 5.86 for exon, and 13.2 for first exon. The same trend was also found for chromosome 22 and after the application of a higher threshold (method 2) on both data sets. In sum, genome-wide analysis has extended earlier analysis and suggests a strong correlation between CpG islands and first coding exons.

4.4 Genome-wide repetitive elements

The proportion of the genome covered by various classes of repetitive DNA is presented in Table 14. We observed about 35% of the genome in these repeat classes, very similar to values reported previously (*83*). Repetitive sequence may be underrepresented in the Celera assembly as a result of incomplete repeat resolution, as discussed above. About 8% of the scaffold length is in gaps, and we expect that much of this is repetitive sequence. Chromosome 19 has the highest repeat density (57%), as well as the highest gene density (Table 10). Of interest, among the different classes of repeat elements, we observe a clear association of Alu elements and gene density, which was not observed between LINEs and gene density.

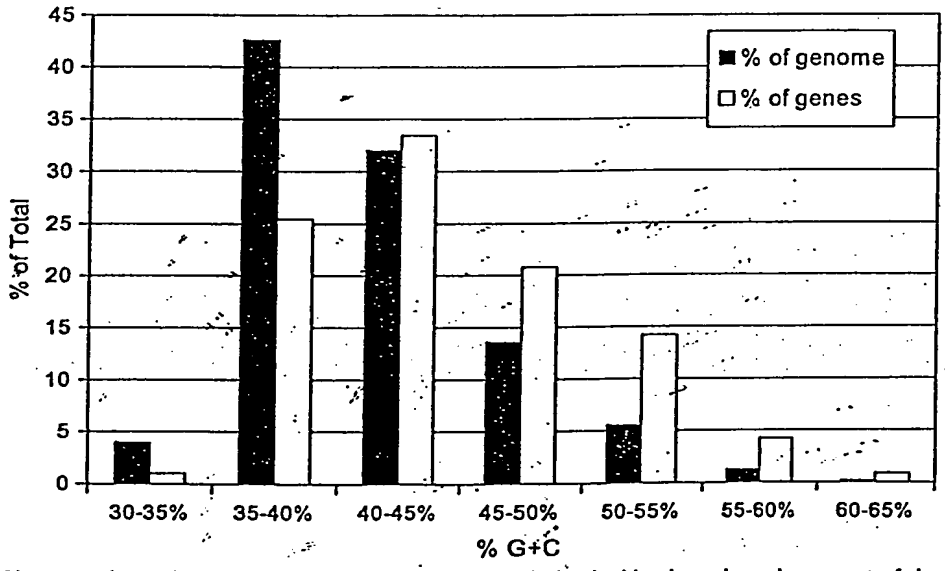


Fig. 10. Relation between G+C content and gene density. The blue bars show the percent of the genome (in 50-kbp windows) with the indicated G+C content. The percent of the total number of genes associated with each G+C bin is represented by the yellow bars. The graph shows that about 5% of the genome has a G+C content of between 50 and 55%, but that this portion contains nearly 15% of the genes.

5 Genome Evolution

Summary. The dynamic nature of genome evolution can be captured at several levels. These include gene duplications mediated by RNA intermediates (retrotransposition) and segmental genomic duplications. In this section, we document the genome-wide occurrence of retrotransposition events generating functional (intronless paralogs) or inactive genes (pseudogenes). Genes involved in translational processes and nuclear regulation account for nearly 50% of all intronless paralogs and processed pseudogenes detected in our survey. We have also cataloged the extent of segmental genomic duplication and provide evidence for 1077 duplicated blocks covering 3522 distinct genes.

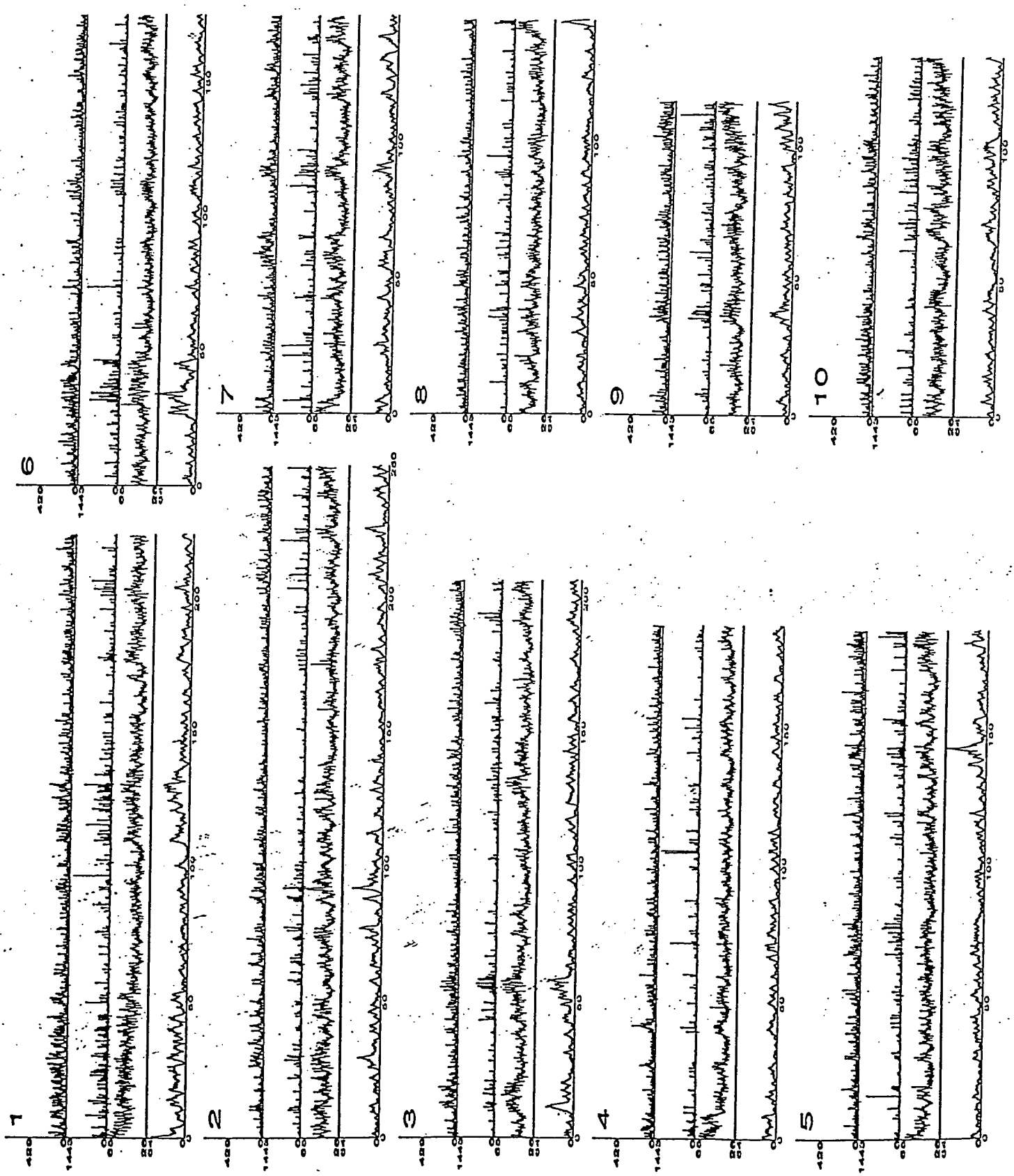


Fig. 11. Genome structural features.

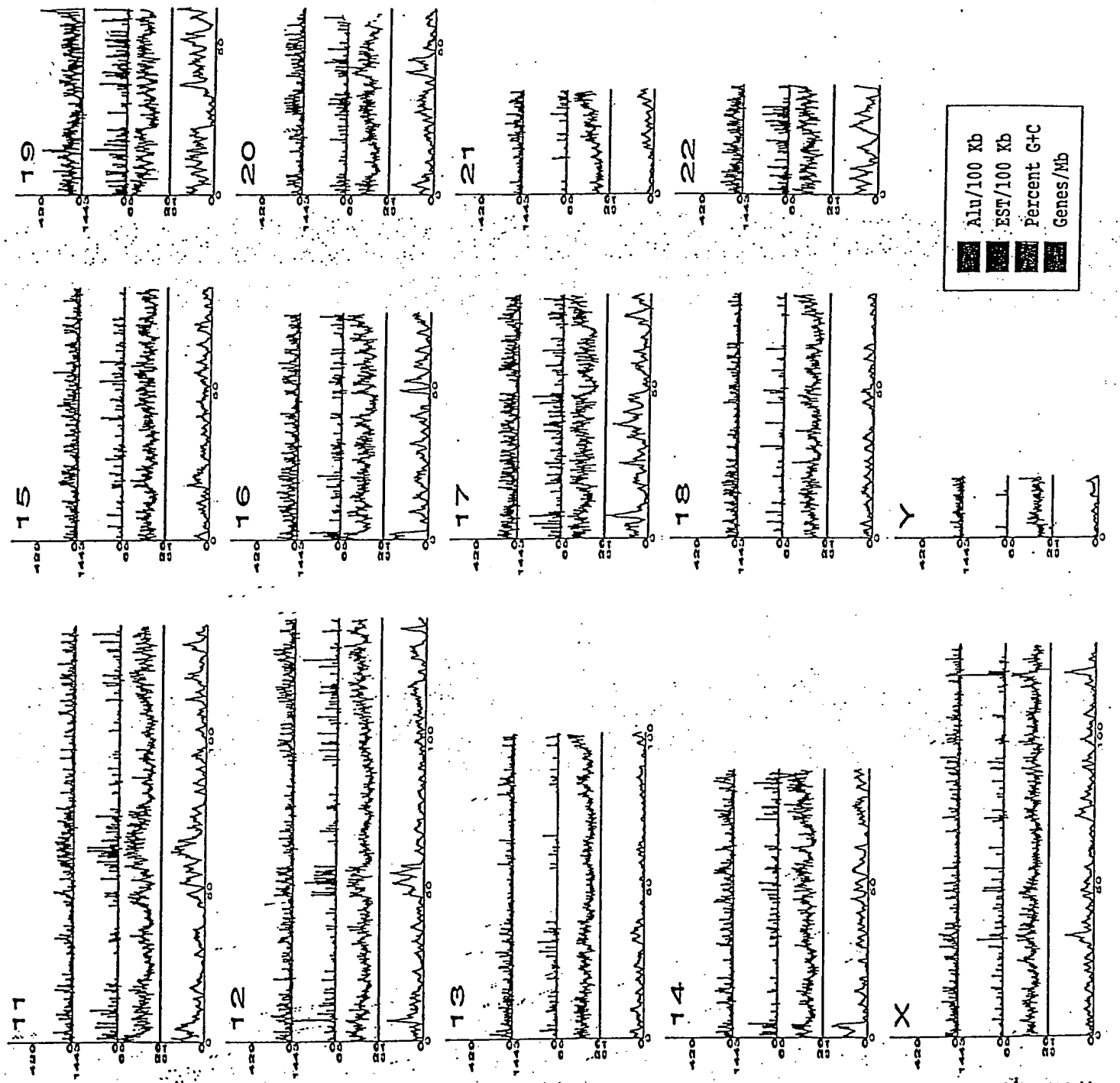


Fig. 11 (continued). Relation among gene density (orange), G+C content (green), EST density (blue), and Alu density (pink) along the lengths of each of the chromosomes. Gene density was calculated in 1-Mbp windows. The percent of G+C nucleotides was calculated in 100-kbp windows. The number of ESTs and Alu elements is shown per 100-kbp window.

5.1 Retrotransposition in the human genome

Retrotransposition of processed mRNA transcripts into the genome results in functional genes, called intronless paralogs, or inactivated genes (pseudogenes). A paralog refers to a gene that appears in more than one copy in a given organism as a result of a duplication event. The existence of both intron-containing and intronless forms of genes encoding functionally similar or identical proteins has been previously described (*84*, *85*). Cataloging these evolutionary events on the genomic landscape is of value in understanding the functional consequences of such gene-duplication events in cellular biology. Identification of conserved intronless paralogs in the mouse or other mammalian genomes should provide the basis for capturing the evolutionary chronology of these transposition events and provide insights into gene loss and accretion in the mammalian radiation.

A set of proteins corresponding to all 901

Table 10. Features of the chromosomes. De novo/any refers to the union of de novo predictions that do not overlap Otto predictions and have at least one other type of supporting evidence; de novo/2x refers to the union of de novo predictions that do not overlap Otto predictions and have at least two types of evidence. Deserts are regions of sequence with no annotated genes.

	Sequence coverage (CS assembly)						Base composition			Gene prediction*					Gene density (genes/Mbp)						
Chr.	Size (Mbp)	No. of scaffolds	Largest scaffold (Mbp)	No. of scaffolds >500 kbp	Sequence covered by scaffolds >500 kbp	% of total sequence in scaffolds >500 kbp	% repeat	% GC	No of CpG Islands	Otto	De novo/ any	De novo/ 2X	Total (Otto + de novo/ any)	Total (Otto + de novo/ any)	Sequence in deserts >500/ kbp	Sequence in deserts >1 Mbp	Otto	De novo/ any	De novo/ 2X	Otto + de novo/ any	Otto + de novo/ 2X
1	220	2,549	11	82	192	88	37	42	2,335	1,743	1,710	710	3,453	2,453	29	6	8	8	3	16	11
2	240	3,263	13	78	217	91	36	40	1,703	1,183	1,771	633	2,954	1,816	55	19	5	7	2	12	7
3	200	3,532	7	78	173	87	37	40	1,271	1,013	1,414	598	2,427	1,611	50	12	5	7	3	12	8
4	186	2,180	10	70	169	91	37	38	1,081	696	1,165	449	1,861	1,145	55	18	4	6	2	10	6
5	182	3,231	11	63	163	89	37	40	1,302	892	1,244	474	2,136	1,366	46	15	5	7	2	11	7
6	172	1,713	13	58	160	93	37	40	1,384	943	1,314	524	2,257	1,467	38	9	6	7	3	13	8
7	146	1,326	14	53	130	89	38	40	1,406	759	1,072	460	1,831	1,219	26	12	5	7	3	12	8
8	146	1,772	11	54	135	92	36	40	948	583	977	357	1,560	940	33	6	4	7	2	11	6
9	113	1,616	8	40	101	89	38	41	1,315	689	848	329	1,537	1,018	22	9	6	7	3	13	8
10	130	2,005	9	55	116	89	36	42	1,087	685	968	342	1,653	1,027	21	8	5	7	2	12	7
11	132	2,814	9	44	116	88	39	42	1,461	1,051	1,134	535	2,185	1,586	27	9	8	8	4	16	12
12	134	2,614	8	51	117	87	38	41	1,131	925	936	417	1,861	1,342	24	9	7	7	3	14	10
13	99	1,038	13	34	91	91	36	38	644	341	691	241	1,032	582	31	16	4	7	2	10	5
14	87	576	11	16	83	95	40	41	913	583	700	290	1,283	873	34	20	7	8	3	14	10
15	80	1,747	8	31	70	87	37	42	722	558	640	246	1,198	804	8	1	7	8	3	15	10
16	75	1,520	8	27	62	82	40	44	1,533	748	673	247	1,421	995	13	3	10	9	3	19	12
17	78	1,683	6	40	61	78	39	45	1,489	897	648	313	1,545	1,210	15	6	12	8	4	19	15
18	79	1,333	13	18	72	92	36	40	510	283	543	189	826	472	21	10	4	7	2	10	6
19	58	2,282	3	31	38	67	57	49	2,804	1,141	534	268	1,675	1,409	3	0	20	9	4	29	23
20	61	580	14	17	58	94	41	44	997	517	469	180	986	697	7	1	8	7	3	16	11
21	33	358	10	6	32	96	38	41	519	184	265	102	449	286	15	9	6	8	3	13	8
22	36	333	11	12	32	88	44	48	1,173	494	341	147	835	641	3	0	14	9	4	23	17
X	128	1,346	4	91	93	73	46	39	726	605	860	387	1,465	992	29	8	5	6	3	11	7
Y	19	638	2	10	12	65	50	39	65	55	155	49	210	104	4	2	3	8	2	11	5
U*	75	11,542	1						479	196	278	132	474	328							
Total	2907	53,591		1,059	2,490				28,519	17,764	21,350	8,619	39,114	26,383	606	208					
Avg.	116	2,144	9	44	104	87	40	41	1,160	714	812	333	1,526	1,047	25	9	7	7	3	14	9

*Chromosomal assignment unknown.

Otto-predicted, single-exon genes were subjected to BLAST analysis against the proteins encoded by the remaining multiexon predicted transcripts. Using homology criteria of 70% sequence identity over 90% of the length, we identified 298 instances of single- to multi-exon correspondence. Of these 298 sequences, 97 were represented in the GenBank data set of experimentally validated full-length genes at the stringency specified and were verified by manual inspection.

We believe that these 97 cases may represent intronless paralogs (see Web table 1 on *Science* Online at www.sciencemag.org/cgi/content/full/291/5507/1304/DC1) of known genes. Most of these are flanked by direct repeat sequences, although the precise nature of these repeats remains to be determined. All of the cases for which we have high confidence contain polyadenylated [poly(A)] tails characteristic of retrotransposition.

Recent publications describing the phenomenon of functional intronless paralogs speculate that retrotransposition may serve as a mechanism used to escape X-chromosomal inactivation (*84*, *86*). We do not find a bias toward X chromosome origination of these retrotransposed genes; rather, the results show a random chromosome distribution of both the intron-containing and corresponding intronless paralogs. We also have found several cases of retrotransposition from a single source chromosome to multiple target chromosomes. Interesting examples include the retrotransposition of a five exon–containing ribosomal protein L21 gene on chromosome 13 onto chromosomes 1, 3, 4, 7, 10, and 14, respectively. The size of the source genes can also show variability. The largest example is the 31-exon diacylglycerol kinase zeta gene on chromosome 11 that has an intronless paralog on chromosome 13. Regardless of route, retrotransposition with subsequent gene changes in coding or noncoding regions that lead to different functions or expression patterns, represents a key route to providing an enhanced functional repertoire in mammals (*87*).

Our preliminary set of retrotransposed intronless paralogs contains a clear overrepresentation of genes involved in translational processes (40% ribosomal proteins and 10% translation elongation factors) and nuclear regulation (HMG nonhistone proteins, 4%), as well as metabolic and regulatory enzymes. EST matches specific to a subset of intronless paralogs suggest expression of these intronless paralogs. Differences in the upstream regulatory sequences between the source genes and their intronless paralogs could account for differences in tissue-specific gene expression. Defining which, if any, of these processed genes are functionally expressed and translated will require further elucidation and experimental validation.

5.2 Pseudogenes

A pseudogene is a nonfunctional copy that is very similar to a normal gene but that has been altered slightly so that it is not expressed. We developed a method for the preliminary analysis of processed pseudogenes in the human genome as a starting point in elucidating the ongoing evolutionary forces

Table 11. Genome overview.

Size of the genome (including gaps)	2.91 Gbp
Size of the genome (excluding gaps)	2.66 Gbp
Longest contig	1.99 Mbp
Longest scaffold	14.4 Mbp
Percent of A+T in the genome	54
Percent of G+C in the genome	38
Percent of undetermined bases in the genome	9
Most GC-rich 50 kb	Chr. 2 (66%)
Least GC-rich 50 kb	Chr. X (25%)
Percent of genome classified as repeats	35
Number of annotated genes	26,383
Percent of annotated genes with unknown function	42
Number of genes (hypothetical and annotated)	39,114
Percent of hypothetical and annotated genes with unknown function	59
Gene with the most exons	Titin (234 exons)
Average gene size	27 kbp
Most gene-rich chromosome	Chr. 19 (23 genes/Mb)
Least gene-rich chromosomes	Chr. 13 (5 genes/Mb), Chr. Y (5 genes/Mb)
Total size of gene deserts (>500 kb with no annotated genes)	605 Mbp
Percent of base pairs spanned by genes	25.5 to 37.8*
Percent of base pairs spanned by exons	1.1 to 1.4*
Percent of base pairs spanned by introns	24.4 to 36.4*
Percent of base pairs in intergenic DNA	74.5 to 63.6*
Chromosome with highest proportion of DNA in annotated exons	Chr. 19 (9.33)
Chromosome with lowest proportion of DNA in annotated exons	Chr. Y (0.36)
Longest intergenic region (between annotated + hypothetical genes)	Chr. 13 (3,038,416 bp)
Rate of SNP variation	1/1250 bp

*In these ranges, the percentages correspond to the annotated gene set (26, 383 genes) and the hypothetical + annotated gene set (39,114 genes), respectively.

Table 12. Rate of recombination per physical distance (cM/Mb) across the genome. Genethon markers were placed on CSA-mapped assemblies, and then relative physical distances and rates were calculated in 3-Mb windows for each chromosome. NA, not applicable.

Chrom.	Male			Sex-average			Female		
	Max.	Avg.	Min.	Max.	Avg.	Min.	Max.	Avg.	Min.
1	2.60	1.12	0.23	2.81	1.42	0.52	3.39	1.76	0.68
2	2.23	0.78	0.33	2.65	1.12	0.54	3.17	1.40	0.61
3	2.55	0.86	0.23	2.40	1.07	0.42	2.71	1.30	0.33
4	1.66	0.67	0.15	2.06	1.04	0.60	2.50	1.40	0.77
5	2.00	0.67	0.18	1.87	1.08	0.42	2.26	1.43	0.62
6	1.97	0.71	0.28	2.57	1.12	0.37	3.47	1.67	0.64
7	2.34	1.16	0.48	1.67	1.17	0.47	2.27	1.21	0.34
8	1.83	0.73	0.14	2.40	1.05	0.46	3.44	1.36	0.43
9	2.01	0.99	0.53	1.95	1.32	0.77	2.63	1.66	0.82
10	3.73	1.03	0.22	3.05	1.29	0.66	2.84	1.51	0.76
11	1.43	0.72	0.31	2.13	0.99	0.47	3.10	1.32	0.49
12	4.12	0.76	0.26	3.35	1.16	0.49	2.93	1.55	0.59
13	1.60	0.75	0.01	1.87	0.95	0.17	2.49	1.19	0.32
14	3.15	0.98	0.18	2.65	1.30	0.62	3.14	1.63	0.75
15	2.28	0.94	0.34	2.31	1.22	0.42	2.53	1.56	0.54
16	1.83	1.00	0.47	2.70	1.55	0.63	4.99	2.32	1.12
17	3.87	0.87	0.00	3.54	1.35	0.54	4.19	1.83	0.94
18	3.12	1.37	0.86	3.75	1.66	0.43	4.35	2.24	0.72
19	3.02	0.97	0.10	2.57	1.41	0.49	2.89	1.75	0.87
20	3.64	0.89	0.00	2.79	1.50	0.83	3.31	2.15	1.34
21	3.23	1.26	0.69	2.37	1.62	1.08	2.58	1.90	1.18
22	1.25	1.10	0.84	1.88	1.41	1.08	3.73	2.08	0.93
X	NA	NA	NA	NA	NA	NA	3.12	1.64	0.72
Y	NA	NA	NA	NA	NA	NA	NA	NA	NA
Genome	4.12	0.88	0.00	3.75	1.22	0.17	4.99	1.55	0.32

that account for gene inactivation. The general structural characteristics of these processed pseudogenes include the complete lack of intervening sequences found in the functional counterparts, a poly(A) tract at the 3′ end, and direct repeats flanking the pseudogene sequence. Processed pseudogenes occur as a result of retrotransposition, whereas unprocessed pseudogenes arise from segmental genome duplication.

We searched the complete set of Otto-predicted transcripts against the genomic sequence by means of BLAST. Genomic regions corresponding to all Otto-predicted transcripts were excluded from this analysis. We identified 2909 regions matching with greater than 70% identity over at least 70% of the length of the transcripts that likely represent processed pseudogenes. This number is probably an underestimate because specific methods to search for pseudogenes were not used.

We looked for correlations between structural elements and the propensity for retrotransposition in the human genome. GC content and transcript length were compared between the genes with processed pseudogenes (1177 source genes) versus the remainder of the predicted gene set. Transcripts that give rise to processed pseudogenes have shorter average transcript length (1027 bp versus 1594 bp for the Otto set) as compared with genes for which no pseudogene was detected. The overall GC content did not show any significant difference, contrary to a recent report (*88*). There is a clear trend in gene families that are present as processed pseudogenes. These include ribosomal proteins (67%), lamin receptors (10%), translation elongation factor alpha (5%), and HMG–non-histone proteins (2%). The increased occurrence of retrotransposition (both intronless paralogs and processed pseudogenes) among genes involved in translation and nuclear regulation may reflect an increased transcriptional activity of these genes.

5.3 Gene duplication in the human genome

Building on a previously published procedure (*27*), we developed a graph-theoretic algorithm, called Lek, for grouping the predicted human protein set into protein families (*89*). The complete clusters that result from the Lek clustering provide one basis for comparing the role of whole-genome or chromosomal duplication in protein family expansion as opposed to other means, such as tandem duplication. Because each complete cluster represents a closed and certain island of homology, and because Lek is capable of simultaneously clustering protein complements of several organisms, the number of proteins contributed by each organism to a complete cluster can be predicted with confidence depending on the quality of the annotation of each genome. The variance of each organism's contribution to each cluster can then be calculated, allowing an assessment of the relative importance of large-scale duplication versus smaller-scale, organism-specific expansion and contraction of protein families, presumably as a result of natural selection operating on individual protein families within an organism. As can be seen in Fig. 12, the large variance in the relative numbers of human as compared with *D. melanogaster* and *Caenorhabditis elegans* proteins in complete clusters may be explained by multiple events of relative expansions in gene families in each of the three animal genomes. Such expansions would give rise to the distribution that shows a peak at 1:1 in the ratio for human-worm or human-fly clusters with the slope spread covering both human and fly/worm predominance, as we observed (Fig. 12). Furthermore, there are nearly as many clusters where worm and fly proteins predominate despite the larger numbers of proteins in the human. At face value, this analysis suggests that natural selection acting on individual protein families has been a major force driving the expansion of at least some elements of the human protein set. However, in our analysis, the difference between an ancient whole-genome duplication followed by loss, versus piecemeal duplication, cannot be easily distinguished. In order to differentiate these scenarios, more extended analyses were performed.

5.4 Large-scale duplications

Using two independent methods, we searched for large-scale duplications in the human genome. First, we describe a protein family–based method that identified highly conserved blocks of duplication. We then describe our comprehensive method for identifying all interchromosomal block duplications. The latter method identified a large number of duplicated chromosomal segments covering parts of all 24 chromosomes.

The first of the methods is based on the idea of searching for blocks of highly conserved homologous proteins that occur in more than one location on the genome. For this comparison, two genes were considered equivalent if their protein products were de-

Table 13. Characteristics of CpG islands identified in chromosome 22 (34-Mbp sequence length) and the whole genome (2.9-Gbp sequence length) by means of two different methods. Method 1 uses a CG likelihood ratio of ≥0.6. Method 2 uses a CG likelihood ratio of ≥0.8.

	Chromosome 22		Whole genome (CS assembly)	
	Method 1	Method 2	Method 1	Method 2
Number of CpG islands detected	5,211	522	195,706	26,876
Average length of island (bp)	390	535	395	497
Percent of sequence predicted as CpG	5.9	0.8	2.6	0.4
Percent of first exons that overlap a CpG island	44	25	42	22
Percent of first exons with first position of exon contained inside a CpG island	37	22	40	21
Average distance between first exon and closest CpG island (bp)	1,013	10,486	2,182	17,021
Expected distance between first exon and closest CpG island (bp)	3,262	32,567	7,164	55,811

Table 14. Distribution of repetitive DNA in the compartmentalized shotgun assembly sequence.

Repetitive elements	Megabases in assembled sequences	Percent of assembly	Previously predicted (%) (*83*)
Alu	288	9.9	10.0
Mammalian interspersed repeat (MIR)	66	2.3	1.7
Medium reiteration (MER)	50	1.7	1.6
Long terminal repeat (LTR)	155	5.3	5.6
Long interspersed nucleotide element (LINE)	466	16.1	16.7
Total	1025	35.3	35.6

termined to be in the same family and the same complete Lek cluster (essentially paralogous genes) (*89*). Initially, each chromosome was represented as a string of genes ordered by the start codons for predicted genes along the chromosome. We considered the two strands as a single string, because local inversions are relatively common events relative to large-scale duplications. Each gene was indexed according to the protein family and Lek complete cluster (*89*). All pairs of indexed gene strings were then aligned in both the forward and reverse directions with the Smith-Waterman algorithm (*90*). A match between two proteins of the same Lek complete cluster was given a score of 10 and a mismatch −10, with gap open and extend penalties of −4 and −1. With these parameters, 19 conserved interchromosomal blocks of duplication were observed, all of which were also detected and expanded by the comprehensive method described below. The detection of only a relatively small number of block duplications was a consequence of using an intrinsically conservative method grounded in the conservative constraints of the complete Lek clusters.

In the second, more comprehensive approach, we aligned all chromosomes directly with one another using an algorithm based on the MUMmer system (*91*). This alignment method uses a suffix tree data structure and a linear-time algorithm to align long sequences very rapidly; for example, two chromosomes of 100 Mbp can be aligned in less than 20 min (on a Compaq Alpha computer) with 4 gigabytes of memory. This procedure was used recently to identify numerous large-scale segmental duplications among the five chromosomes of *A. thaliana* (*92*); in that organism, the method revealed that 60% of the genome (66 Mbp) is covered by 24 very large duplicated segments. For *Arabidopsis*, a DNA-based alignment was sufficient to reveal the segmental duplications between chromosomes; in the human genome, DNA alignments at the whole-chromosome level are insufficiently sensitive. Therefore, a modified procedure was developed and applied, as follows. First, all 26,588 proteins (9,675,713 million amino acids) were concatenated end-to-end in order as they occur along each of the 24 chromosomes, irrespective of strand location. The concatenated protein set was then aligned against each chromosome by the MUMmer algorithm. The resulting matches were clustered to extract all sets of three or more protein matches that occur in close proximity on two different chromosomes (*93*); these represent the candidate segmental duplications. A series of filters were developed and applied to remove likely false-positives from this set; for example, small blocks that were spread across many proteins were removed. To refine the filtering methods, a shuffled protein set was first created by taking the 26,588 proteins, randomizing their order, and then partitioning them into 24 shuffled chromosomes, each containing the same number of proteins as the true genome. This shuffled protein set has the identical composition to the real genome; in particular, every protein and every domain appears the same number of times. The complete algorithm was then applied to both the real and the shuffled data, with the results on the shuffled data being used to estimate the false-positive rate. The algorithm after filtering yielded 10,310 gene pairs in 1077 duplicated blocks containing 3522 distinct genes; tandemly duplicated expansions in many of the blocks explain the excess of gene pairs to distinct genes. In the shuffled data, by contrast, only 370 gene pairs were found, giving a false-positive estimate of 3.6%. The most likely explanation for the 1077 block duplications is ancient segmental duplications. In many cases, the order of the proteins has been shuffled, although proximity is preserved. Out of the 1077 blocks, 159 contain only three genes, 137 contain four genes, and 781 contain five or more genes.

To illustrate the extent of the detected duplications, Fig. 13 shows all 1077 block duplications indexed to each chromosome in 24 panels in which only duplications mapped to the indexed chromosome are displayed. The figure makes it clear that the duplications are ubiquitous in the genome. One feature that it displays is many relatively small chromosomal stretches, with one-to-many duplication relationships that are graphically striking. One such example captured by the analysis is the well-documented olfactory receptor (OR) family, which is scattered in blocks throughout the genome and which has been analyzed for genome-deployment reconstructions at several evolutionary stages (*94*). The figure also illustrates that some chromosomes, such as chromosome 2, contain many more detected large-scale duplications than others. Indeed, one of the largest duplicated segments is a large block of 33 proteins on chromosome 2, spread among eight smaller blocks in 2p, that aligns to a paralogous set on chromosome 14, with one rearrangement (see chromosomes 2 and 14 panels in Fig. 13). The proteins are not contiguous but span a region containing 97 proteins on chromosome 2 and 332 proteins on chromosome 14. The likelihood of observing this many duplicated proteins by chance, even over a span of this length, is 2.3×10^{-68} (*93*). This duplicated set spans 20 Mbp on chromosome 2 and 63 Mbp on chromosome 14, over 70% of the latter chromosome. Chromosome 2 also contains a block duplication that is nearly as large, which is shared by chromosome arm 2q and chromosome 12. This duplication incorporates two of the four known Hox gene clusters, but considerably expands the extent of the duplications proximally and distally on the pair of chromosome arms. This breadth of duplication is also seen on the two chromosomes carrying the other two Hox clusters.

An additional large duplication, between chromosomes 18 and 20, serves as a good example to illustrate some of the features common to many of the other observed large duplications (Fig. 13, inset). This duplication contains 64 detected ordered intrachromosomal pairs of homologous genes. After discounting a 40-Mb stretch of chromosome 18 free of matches to chromosome 20, which is likely to represent a large insert (between the gene assignments "Krup rel" and "collagen rel" on chromosome 18 in Fig. 13), the full duplication segment covers 36 Mb on chromosome 18 and 28 Mb on chromosome 20.

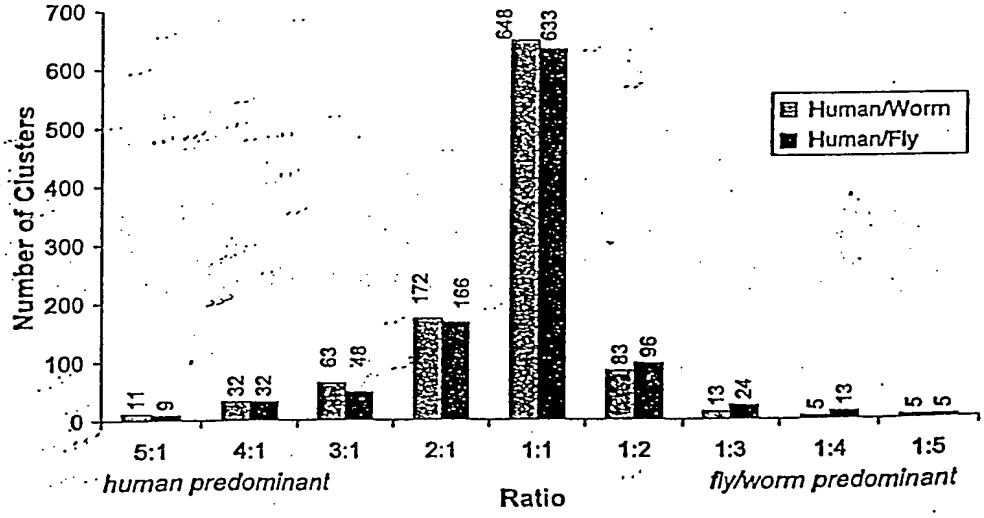


Fig. 12. Gene duplication in complete protein clusters. The predicted protein sets of human, worm, and fly were subjected to Lek clustering (27). The numbers of clusters with varying ratios (whole number) of human versus worm and human versus fly proteins per cluster were plotted.

By this measure, the duplication segment spans nearly half of each chromosome's net length. The most likely scenario is that the whole span of this region was duplicated as a single very large block, followed by shuffling owing to smaller scale rearrangements. As such, at least four subsequent rearrangements would need to be invoked to explain the relative insertions and inversions seen in the duplicated segment interval. The 64 protein pairs in this alignment occur among 217 protein assignments on chromosome 18, and among 322 protein assignments on chromosome 20, for a density of involved proteins of 20 to 30%. This is consistent with an ancient large-scale duplication followed by subsequent gene loss on one or both chromosomes. Loss of just one member of a gene pair subsequent to the duplication would result in a failure to score a gene pair in the block; less than 50% gene loss on the chromosomes would lead to the duplication density observed here. As an independent verification of the significance of the alignments detected, it can be seen that a substantial number of the pairs of aligning proteins in this duplication, including some of those annotated (Fig. 13), are those populating small Lek complete clusters (see above). This indicates that they are members of very small families of paralogs; their relative scarcity within the genome validates the uniqueness and robust nature of their alignments.

Two additional qualitative features were observed among many of the large-scale duplications. First, several proteins with disease associations, with OMIM (Online Mendelian Inheritance in Man) assignments, are members of duplicated segments (see web table 2 on *Science* Online at www.sciencemag.org/cgi/content/full/291/5507/1304/DC1). We have also observed a few instances where paralogs on both duplicated segments are associated with similar disease conditions. Notable among these genes are proteins involved in hemostasis (coagulation factors) that are associated with bleeding disorders, transcriptional regulators like the homeobox proteins associated with developmental disorders, and potassium channels associated with cardiovascular conduction abnormalities. For each of these disease genes, closer study of the paralogous genes in the duplicated segment may reveal new insights into disease causation, with further investigation needed to determine whether they might be involved in the same or similar genetic diseases. Second, although there is a conserved number of proteins and coding exons predicted for specific large duplicated spans within the chromosome 18 to 20 alignment, the genomic DNA of chromosome 18 in these specific spans is in some cases more than 10-fold longer than the corresponding chromosome 20 DNA. This selective accretion of noncoding DNA (or conversely, loss of noncoding DNA) on one of a pair of duplicated chromosome regions was observed in many compared regions. Hypotheses to explain which mechanisms foster these processes must be tested.

Evaluation of the alignment results gives some perspective on dating of the duplications. As noted above, large-scale ancient segmental duplication in fact best explains many of the blocks detected by this genome-wide analysis. The regions of human chromosomes involved in the large-scale duplications expanded upon above (chromosomes 2 to 14, 2 to 12, and 18 to 20) are each syntenic to a distinct mouse chromosomal region. The corresponding mouse chromosomal regions are much more similar in sequence conservation, and even in order, to their human synteny partners than the human duplication regions are to each other. Further, the corresponding mouse chromosomal regions each bear a significant proportion of genes orthologous to the human genes on which the human duplication assignments were made. On the basis of these factors, the corresponding mouse chromosomal spans, at coarse resolution, appear to be products of the same large-scale duplications observed in humans. Although further detailed analysis must be carried out once a more complete genome is assembled for mouse, the underlying large duplications appear to predate the two species' divergence. This dates the duplications, at the latest, before divergence of the primate and rodent lineages. This date can be further refined upon examination of the synteny between human chromosomes and those of chicken, pufferfish (*Fugu rubripes*), or zebrafish (*95*). The only substantial syntenic stretches mapped in these species corresponding to both pairs of human duplications are restricted to the Hox cluster regions. When the synteny of these regions (or others) to human chromosomes is extended with further mapping, the ages of the nearly chromosome-length duplications seen in humans are likely to be dated to the root of vertebrate divergence.

The MUMmer-based results demonstrate large block duplications that range in size from a few genes to segments covering most of a chromosome. The extent of segmental duplications raises the question of whether an ancient whole-genome duplication event is the underlying explanation for the numerous duplicated regions (*96*). The duplications have undergone many deletions and subsequent rearrangements; these events make it difficult to distinguish between a whole-genome duplication and multiple smaller events. Further analysis, focused especially on comparing the estimated ages of all the block duplications, derived partially from interspecies genome comparisons, will be necessary to determine which of these two hypotheses is more likely. Comparisons of genomes of different vertebrates, and even cross-phyla genome comparisons, will allow for the deconvolution of duplications to eventually reveal the stagewise history of our genome, and with it a history of the emergence of many of the key functions that distinguish us from other living things.

6 A Genome-Wide Examination of Sequence Variations

Summary. Computational methods were used to identify single-nucleotide polymorphisms (SNPs) by comparison of the Celera sequence to other SNP resources. The SNP rate between two chromosomes was ~1 per 1200 to 1500 bp. SNPs are distributed nonrandomly throughout the genome. Only a very small proportion of all SNPs (<1%) potentially impact protein function based on the functional analysis of SNPs that affect the predicted coding regions. This results in an estimate that only thousands, not millions, of genetic variations may contribute to the structural diversity of human proteins.

Having a complete genome sequence enables researchers to achieve a dramatic acceleration in the rate of gene discovery, but only through analysis of sequence variation in DNA can we discover the genetic basis for variation in health among human beings. Whole-genome shotgun sequencing is a particularly effective method for detecting sequence variation in tandem with whole-genome assembly. In addition, we compared the distribution and attributes of SNPs ascertained by three other methods: (i) alignment of the Celera consensus sequence to the PFP assembly, (ii) overlap of high-quality reads of genomic sequence (referred to as "Kwok"; 1,120,195 SNPs) (*97*), and (iii) reduced representation shotgun sequencing (referred to as "TSC"; 632,640 SNPs) (*98*). These data were consistent in showing an overall nucleotide diversity of ~8×10^{-4}, marked heterogeneity across the genome in SNP density, and an overwhelming preponderance of noncoding variation that produces no change in expressed proteins.

6.1 SNPs found by aligning the Celera consensus to the PFP assembly

Ideally, methods of SNP discovery make full use of sequence depth and quality at every site, and quantitatively control the rate of false-positive and false-negative calls with an explicit sampling model (*99*). Comparison of consensus sequences in the absence of these details necessitated a more ad hoc approach (quality scores could not readily be obtained for the PFP assembly). First, all sequence differences between the two consensus sequences were identified; these were then filtered to reduce the contribution of sequencing errors and misassembly. As a measure of the effectiveness of the filtering step, we monitored the ratio of transition and transversion substitutions, because a 2:1 ratio has been well documented as typical in mammalian evolution (*100*) and in human SNPs

(*101*, *102*). The filtering steps consisted of removing variants where the quality score in the Celera consensus was less than 30 and where the density of variants was greater than 5 in 400 bp. These filters resulted in shifting the transition-to-transversion ratio from 1.57:1 to 1.89:1. When applied to 2.3 Gbp of alignments between the Celera and PFP consensus sequences, these filters resulted in identification of 2,104,820 putative SNPs from a total of 2,778,474 substitution differences. Overlaps between this set of SNPs and those found by other methods are described below.

6.2 Comparisons to public SNP databases

Additional SNPs, including 2,536,021 from dbSNP (www.ncbi.nlm.nih.gov/SNP) and 13,150 from HGMD (Human Gene Mutation Database, from the University of Wales, UK), were mapped on the Celera consensus sequence by a sequence similarity search with the program PowerBlast (*103*). The two largest data sets in dbSNP are the Kwok and TSC sets, with 47% and 25% of the dbSNP records. Low-quality alignments with partial coverage of the dbSNP sequence and alignments that had less than 98% sequence identity between the Celera sequence and the dbSNP flanking sequence were eliminated. dbSNP sequences mapping to multiple locations on the Celera genome were discarded. A total of 2,336,935 dbSNP variants were mapped to 1,223,038 unique locations on the Celera sequence, implying considerable redundancy in dbSNP. SNPs in the TSC set mapped to 585,811 unique genomic locations, and SNPs in the Kwok set mapped to 438,032 unique locations. The combined unique SNPs counts used in this analysis, including Celera-PFP, TSC, and Kwok, is 2,737,668. Table 15 shows that a substantial fraction of SNPs identified by one of these methods was also found by another method. The very high overlap (36.2%) between the Kwok and Celera-PFP SNPs may be due in part to the use by Kwok of sequences that went into the PFP assembly. The unusually low overlap (16.4%) between the Kwok and TSC sets is due to their being the smallest two sets. In addition, 24.5% of the Celera-PFP SNPs overlap with SNPs derived from the Celera genome sequences (*46*). SNP validation in population samples is an expensive and laborious process, so confirmation on multiple data sets may provide an efficient initial validation "in silico" (by computational analysis).

One means of assessing whether the three sets of SNPs provide the same picture of human variation is to tally the frequencies of the six possible base changes in each set of SNPs (Table 16). Previous measures of nucleotide diversity were mostly derived from small-scale analysis on candidate genes (*101*), and our analysis with all three data sets validates the previous observations at the whole-genome scale. There is remarkable homogeneity between the SNPs found in the Kwok set, the TSC set, and in our whole-genome shotgun (*46*) in this substitution pattern. Compared with the rest of the data sets, Celera-PFP deviates slightly from the 2:1 transition-to-transversion ratio observed in the other SNP sets. This result is not unexpected, because some fraction of the computationally identified SNPs in the Celera-PFP comparison may in fact be sequence errors. A 2:1 transition:transversion ratio for the bona fide SNPs would be obtained if one assumed that 15% of the sequence differences in the Celera-PFP set were a result of (presumably random) sequence errors.

6.3 Estimation of nucleotide diversity from ascertained SNPs

The number of SNPs identified varied widely across chromosomes. In order to normalize these values to the chromosome size and sequence coverage, we used π, the standard statistic for nucleotide diversity (*104*). Nucleotide diversity is a measure of per-site heterozygosity, quantifying the probability that a pair of chromosomes drawn from the population will differ at a nucleotide site. In order to calculate nucleotide diversity for each chromosome, we need to know the number of nucleotide sites that were surveyed for variation, and in methods like reduced representation sequencing, we need to know the sequence quality and the depth of coverage at each site. These data are not readily available, so we could not estimate nucleotide diversity from the TSC effort. Estimation of nucleotide diversity from high-quality sequence overlaps should be possible, but again, more information is needed on the details of all the alignments.

Estimation of nucleotide diversity from a shotgun assembly entails calculating for each column of the multialignment, the probability that two or more distinct alleles are present, and the probability of detecting a SNP if in fact the alleles have different sequence (i.e., the probability of correct sequence calls). The greater the depth of coverage and the higher the sequence quality, the higher is the chance of successfully detecting a SNP (*105*). Even after correcting for variation in coverage, the nucleotide diversity appeared to vary across autosomes. The significance of this heterogeneity was tested by analysis of variance, with estimates of π for 100-kbp windows to estimate variability within chromosomes (for the Celera-PFP comparison, $F = 29.73$, $P < 0.0001$).

Average diversity for the autosomes estimated from the Celera-PFP comparison was 8.94×10^{-4}. Nucleotide diversity on the X chromosome was 6.54×10^{-4}. The X is expected to be less variable than autosomes, because for every four copies of autosomes in the population, there are only three X chromosomes, and this smaller effective population size means that random drift will more rapidly remove variation from the X (*106*).

Having ascertained nucleotide variation genome-wide, it appears that previous estimates of nucleotide diversity in humans based on samples of genes were reasonably accurate (*101*, *102*, *106*, *107*). Genome-wide, our estimate of nucleotide diversity was 8.98×10^{-4} for the Celera-PFP alignment, and a published estimate averaged over 10 densely resequenced human genes was 8.00×10^{-4} (*108*).

6.4 Variation in nucleotide diversity across the human genome

Such an apparently high degree of variability among chromosomes in SNP density raises the question of whether there is heterogeneity at a finer scale within chromo-

able 15. Overlap of SNPs from genome-wide NP databases. Table entries are SNP counts for ach pair of data sets. Numbers in parentheses are ie fraction of overlap, calculated as the count of verlapping SNPs divided by the number of SNPs i the smaller of the two databases compared. otal SNP counts for the databases are: Celera-P, 2,104,820; TSC, 585,811; and Kwok 438,032. nly unique SNPs in the TSC and Kwok data sets ere included.

	TSC	Kwok
lera-PFP	188,694 (0.322)	158,532 (0.362)
C		72,024 (0.164)

Table 16. Summary of nucleotide changes in different SNP data sets.

SNP data set	A/G (%)	C/T (%)	A/C (%)	A/T (%)	C/G (%)	T/G (%)	Transition: transversion
Celera-PFP	30.7	30.7	10.3	8.6	9.2	10.3	1.59:1
Kwok*	33.7	33.8	8.5	7.0	8.6	8.4	2.07:1
TSC†	33.3	33.4	8.8	7.3	8.6	8.6	1.99:1

*November 2000 release of the NCBI database dbSNP (www.ncbi.nlm.nih.gov/SNP/) with the method defined as Overlap SnpDetectionWithPolyBayes. The submitter of the data is Pui-Yan Kwok from Washington University. †November 2000 release of NCBI dbSNP (www.ncbi.nlm.nih.gov/SNP/) with the methods defined as TSC-Sanger, TSC-WICGR, and TSC-WUGSC. The submitter of the data is Lincoln Stein from Cold Spring Harbor Laboratory.

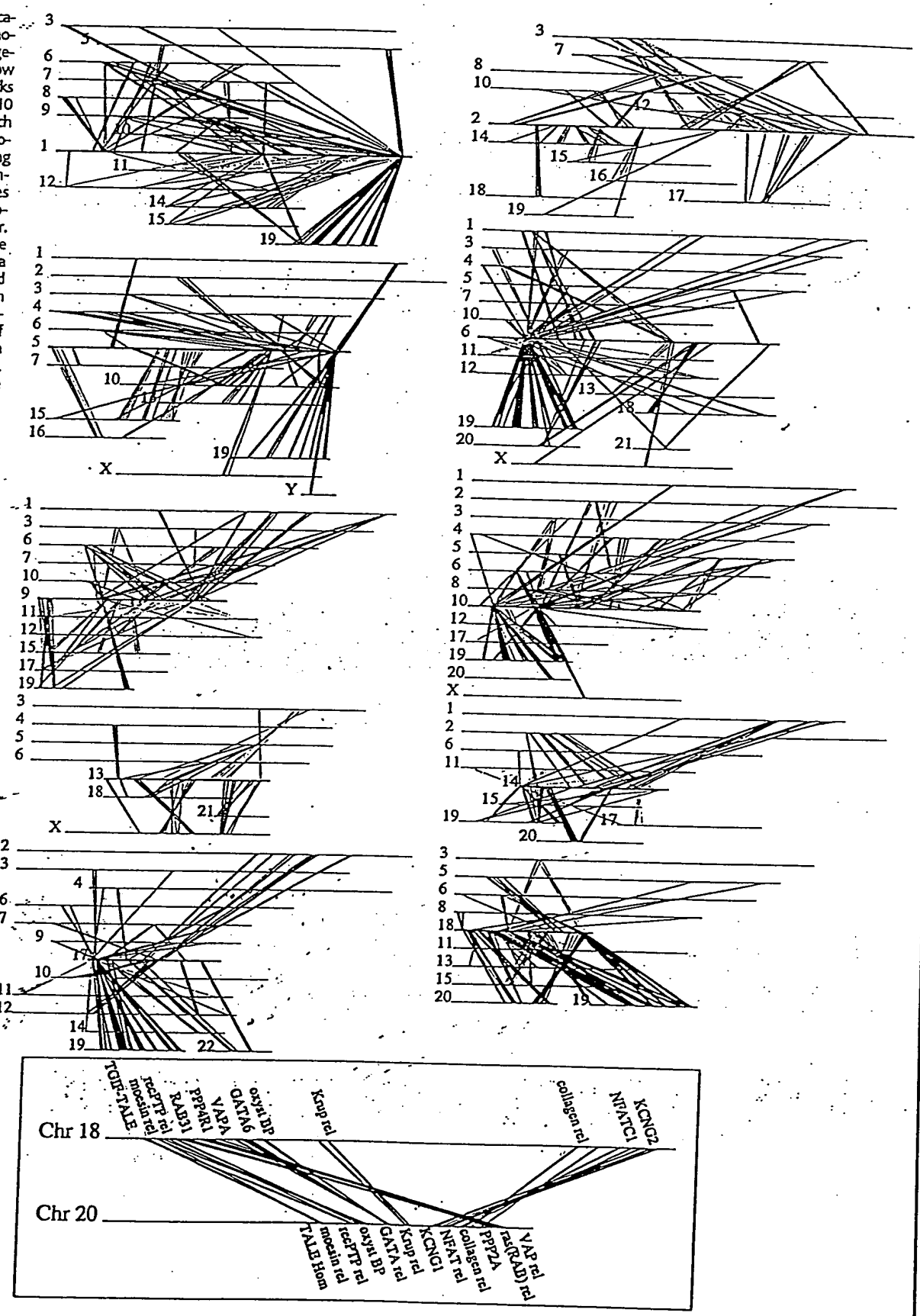
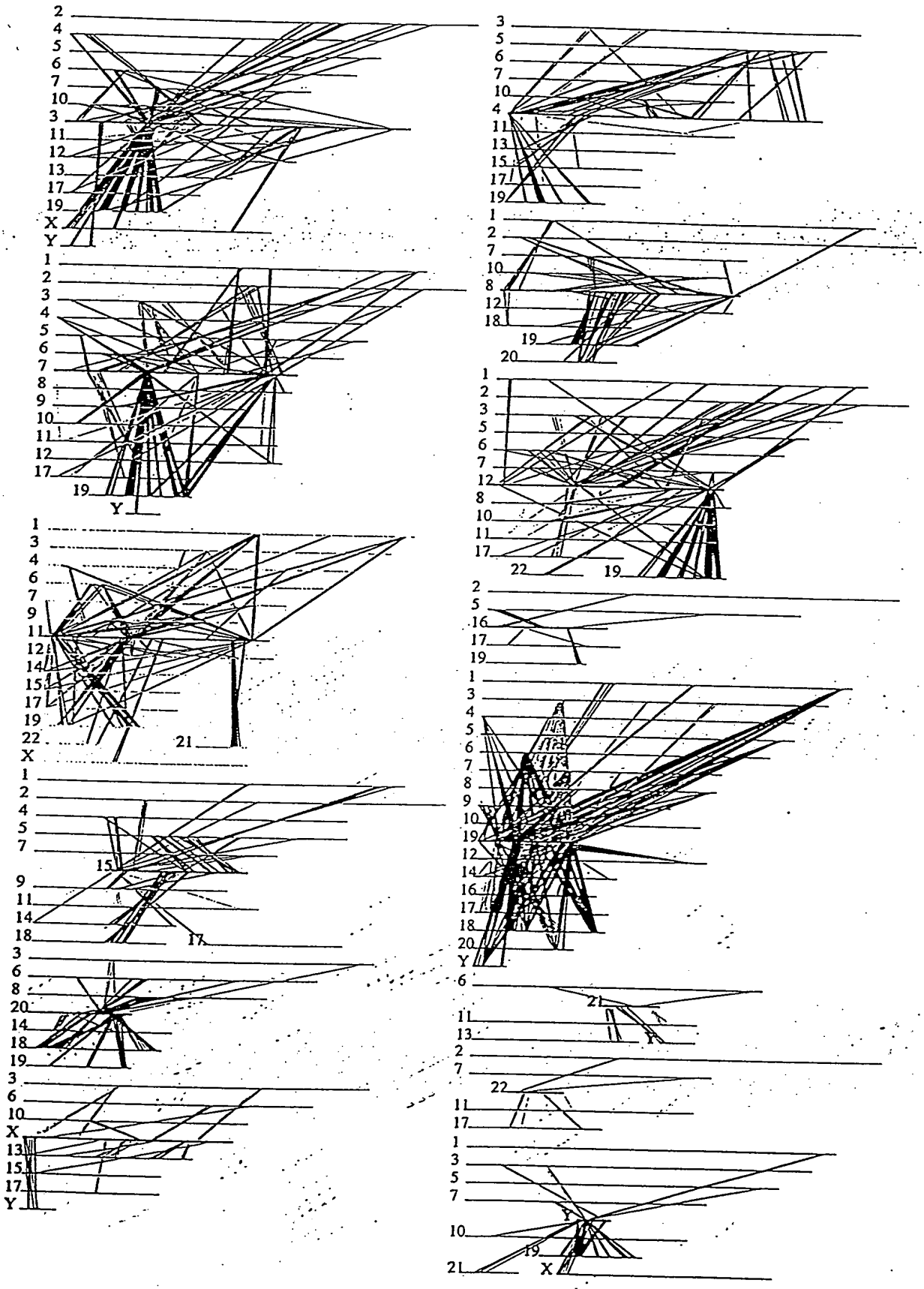


Fig. 13. Segmental duplications between chromosomes in the human genome. The 24 panels show the 1077 duplicated blocks of genes, containing 10,310 pairs of genes in total. Each line represents a pair of homologous genes belonging to a block; all blocks contain at least three genes on each of the chromosomes where they appear. Each panel shows all the duplications between a single chromosome and other chromosomes with shared blocks. The chromosome at the center of each panel is shown as a thick red line for emphasis. Other chromosomes are displayed from top to bottom within each panel ordered by chromosome number. The inset (bottom, center right) shows a close-up of one duplication between chromosomes 18 and 20, expanded to display the gene names of 12 of the 64 gene pairs shown.



somes, and whether this heterogeneity is greater than expected by chance. If SNPs occur by random and independent mutations, then it would seem that there ought to be a Poisson distribution of numbers of SNPs in fragments of arbitrary constant size. The observed dispersion in the distribution of SNPs in 100-kbp fragments was far greater than predicted from a Poisson distribution (Fig. 14). However, this simplistic model ignores the different recombination rates and population histories that exist in different regions of the genome. Population genetics theory holds that we can account for this variation with a mathematical formulation called the neutral coalescent (*109*). Applying well-tested algorithms for simulating the neutral coalescent with recombination (*110*), and using an effective population size of 10,000 and a per-base recombination rate equal to the mutation rate (*111*), we generated a distribution of numbers of SNPs by this model as well (*112*). The observed distribution of SNPs has a much larger variance than either the Poisson model or the coalescent model, and the difference is highly significant. This implies that there is significant variability across the genome in SNP density, an observation that begs an explanation.

Several attributes of the DNA sequence may affect the local density of SNPs, including the rate at which DNA polymerase makes errors and the efficacy of mismatch repair. One key factor that is likely to be associated with SNP density is the G+C content, in part because methylated cytosines in CpG dinucleotides tend to undergo deamination to form thymine, accounting for a nearly 10-fold increase in the mutation rate of CpGs over other dinucleotides. We tallied the GC content and nucleotide diversities in 100-kbp windows across the entire genome and found that the correlation between them was positive ($r = 0.21$) and highly significant ($P < 0.0001$), but G+C content accounted for only a small part of the variation.

6.5 SNPs by genomic class

To test homogeneity of SNP densities across functional classes, we partitioned sites into intergenic (defined as >5 kbp from any predicted transcription unit), 5′-UTR, exonic (missense and silent), intronic, and 3′-UTR for 10,239 known genes, derived from the NCBI RefSeq database and all human genes predicted from the Celera Otto annotation. In coding regions, SNPs were categorized as either silent, for those that do not change amino acid sequence, or missense, for those that change the protein product. The ratio of missense to silent coding SNPs in Celera-PFP, TSC, and Kwok sets (1.12, 0.91, and 0.78, respectively) shows a markedly reduced frequency of missense variants compared with the neutral expectation, consistent with the elimination by natural selection of a fraction of the deleterious amino acid changes (*112*). These ratios are comparable to the missense-to-silent ratios of 0.88 and 1.17 found by Cargill *et al.* (*101*) and by Halushka *et al.* (*102*). Similar results were observed in SNPs derived from Celera shotgun sequences (*46*).

It is striking how small is the fraction of SNPs that lead to potentially dysfunctional alterations in proteins. In the 10,239 RefSeq genes, missense SNPs were only about 0.12, 0.14, and 0.17% of the total SNP counts in Celera-PFP, TSC, and Kwok SNPs, respectively. Nonconservative protein changes constitute an even smaller fraction of missense SNPs (47, 41, and 40% in Celera-PFP, Kwok, and TSC). Intergenic regions have been virtually unstudied (*113*), and we note that 75% of the SNPs we identified were intergenic (Table 17). The SNP rate was highest in introns and lowest in exons. The SNP rate was lower in intergenic regions than in introns, providing one of the first discriminators between these two classes of DNA. These SNP rates were confirmed in the Celera SNPs, which also exhibited a lower rate in exons than in introns, and in extragenic regions than in introns (*46*). Many of these intergenic SNPs will provide valuable information in the form of markers for linkage and association studies, and some fraction is likely to have a regulatory function as well.

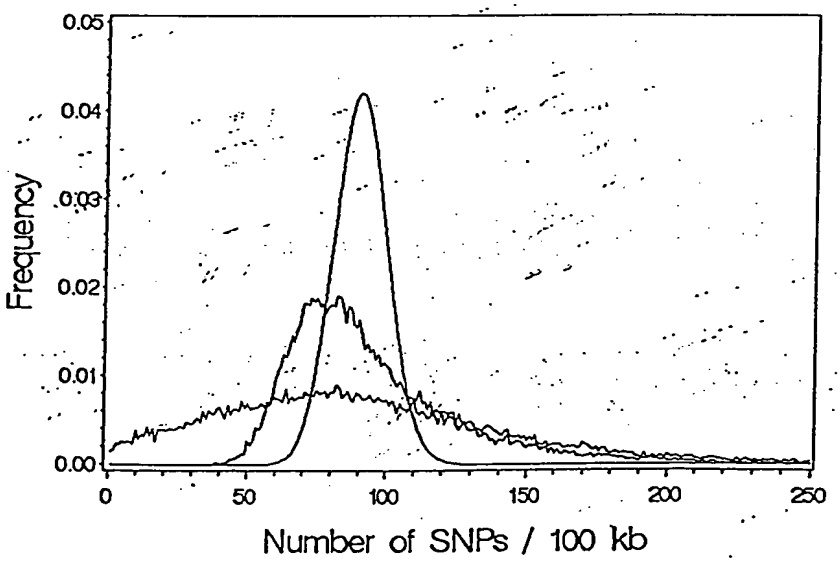


Fig. 14. SNP density in each 100-kbp interval as determined with Celera-PFP SNPs. The color codes are as follows: black, Celera-PFP SNP density; blue, coalescent model; and red, Poisson distribution. The figure shows that the distribution of SNPs along the genome is nonrandom and is not entirely accounted for by a coalescent model of regional history.

7 An Overview of the Predicted Protein-Coding Genes in the Human Genome

Summary. This section provides an initial computational analysis of the predicted protein set with the aim of cataloging prominent differences and similarities when the human genome is compared with other fully sequenced eukaryotic genomes. Over 40% of the predicted protein set in humans cannot be ascribed a molecular function by methods that assign proteins to known families. A protein domain–based analysis provides a detailed catalog of the prominent differences in the human genome when compared with the fly and worm genomes. Prominent among these are domain expansions in proteins involved in developmental regulation and in cellular processes such as neuronal function, hemostasis, acquired immune response, and cytoskeletal complexity. The final enumeration of protein families and details of protein structure will rely on additional experimental work and comprehensive manual curation.

A preliminary analysis of the predicted human protein-coding genes was conducted. Two methods were used to analyze and classify the molecular functions of 26,588 predicted proteins that represent 26,383 gene predictions with at least two lines of evidence as described above. The first method was based on an analysis at the level of protein families, with both the publicly available Pfam database (*114*, *115*) and Celera's Panther Classification (CPC) (Fig. 15) (*116*). The second method was based on an analysis at the level of protein domains, with both the Pfam and SMART databases (*115*, *117*).

The results presented here are preliminary and are subject to several limit-

Both the gene predictions and functional assignments have been made by using computational tools, although the statistical models in Panther, Pfam, and SMART have been built, annotated, and reviewed by expert biologists. In the set of computationally predicted genes, we expect both false-positive predictions (some of these may in fact be inactive pseudogenes) and false-negative predictions (some human genes will not be computationally predicted). We also expect errors in delimiting the boundaries of exons and genes. Similarly, in the automatic functional assignments, we also expect both false-positive and false-negative predictions. The functional assignment protocol focuses on protein families that tend to be found across several organisms, or on families of known human genes. Therefore, we do not assign a function to many genes that are not in large families, even if the function is known. Unless otherwise specified, all enumeration of the genes in any given family or functional category was taken from the set of 26,588 predicted proteins, which were assigned functions by using statistical score cutoffs defined for models in Panther, Pfam, and SMART.

For this initial examination of the predicted human protein set, three broad questions were asked: (i) What are the likely molecular functions of the predicted gene products, and how are these proteins categorized with current classification methods? (ii) What are the core functions that appear to be common across the animals? (iii) How does the human protein complement differ from that of other sequenced eukaryotes?

7.1 Molecular functions of predicted human proteins

Figure 15 shows an overview of the putative molecular functions of the predicted 26,588 human proteins that have at least two lines of supporting evidence. About 41% (12,809) of the gene products could not be classified from this initial analysis and are termed proteins with unknown functions. Because our automatic classification methods treat only relatively large protein families, there are a number of "unclassified" sequences that do, in fact, have a known or predicted function. For the 60% of the protein set that have automatic functional predictions, the specific protein functions have been placed into broad classes. We focus here on molecular function (rather than higher order cellular processes) in order to classify as many proteins as possible. These functional predictions are based on similarity to sequences of known function.

In our analysis of the 12,731 additional low-confidence predicted genes (those with only one piece of supporting evidence), only 636 (5%) of these additional putative genes were assigned molecular functions by the automated methods. One-third of these 636 predicted genes represented endogenous retroviral proteins, further suggesting that the majority of these unknown-function genes are not real genes. Given that most of these additional 12,095 genes appear to be unique among the genomes sequenced to date, many may simply represent false-positive gene predictions.

The most common molecular functions are the transcription factors and those involved in nucleic acid metabolism (nucleic acid enzyme). Other functions that are highly represented in the human genome are the receptors, kinases, and hydrolases. Not surprisingly, most of the hydrolases are proteases. There are also many proteins that are members of proto-oncogene families, as well as families of "select regulatory molecules": (i) proteins involved in specific steps of signal transduction such as heterotrimeric GTP-binding proteins (G proteins) and cell cycle regulators, and (ii) proteins that modulate the activity of kinases, G proteins, and phosphatases.

Table 17. Distribution of SNPs in classes of genomic regions.

Genomic region class	Size of region examined (Mb)	Celera-PFP SNP density (SNP/Mb)
Intergenic	2185	707
Gene (intron + exon)	646	917
Intron	615	921
First intron	164	808
Exon	31	529
First exon	10	592

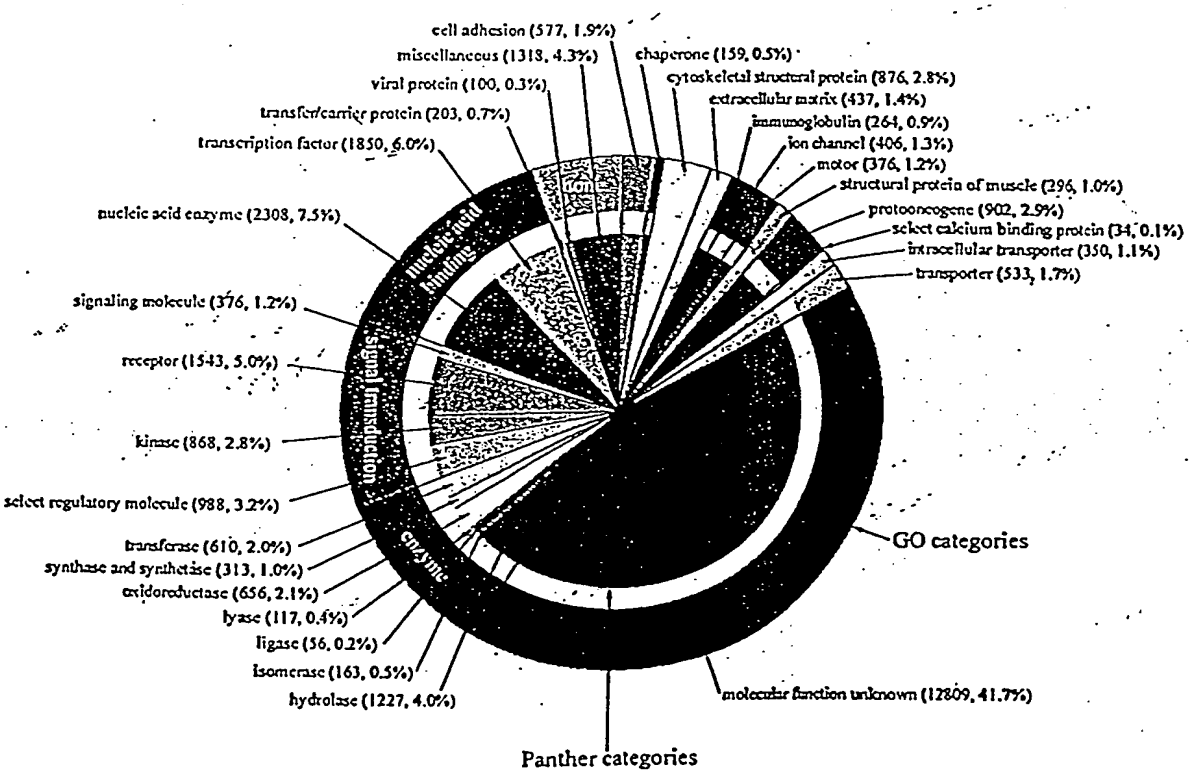


Fig. 15. Distribution of the molecular functions of 26,383 human genes. Each slice lists the numbers and percentages (in parentheses) of human gene functions assigned to a given category of molecular function. The outer circle shows the assignment to molecular function categories in the Gene Ontology (GO) (*179*), and the inner circle shows the assignment to Celera's Panther molecular function categories (*116*).

7.2 Evolutionary conservation of core processes

Because of the various "model organism" genome-sequencing projects that have already been completed, reasonable comparative information is available for beginning the analysis of the evolution of the human genome. The genomes of *S. cerevisiae* ("bakers' yeast") (*118*) and two diverse invertebrates, *C. elegans* (a nematode worm) (*119*) and *D. melanogaster* (fly) (*26*), as well as the first plant genome, *A. thaliana*, recently completed (*92*), provide a diverse background for genome comparisons.

We enumerated the "strict orthologs" conserved between human and fly, and between human and worm (Fig. 16) to address the question, What are the core functions that appear to be common across the animals? The concept of orthology is important because if two genes are orthologs, they can be traced by descent to the common ancestor of the two organisms (an "evolutionarily conserved protein set"), and therefore are likely to perform similar conserved functions in the different organisms. It is critical in this analysis to separate orthologs (a gene that appears in two organisms by descent from a common ancestor) from paralogs (a gene that appears in more than one copy in a given organism by a duplication event) because paralogs may subsequently diverge in function. Following the yeast-worm ortholog comparison in (*120*), we identified two different cases for each pairwise comparison (human-fly and human-worm). The first case was a pair of genes, one from each organism, for which there was no other close homolog in either organism. These are straightforwardly identified as orthologous, because there are no additional members of the families that complicate separating orthologs from paralogs. The second case is a family of genes with more than one member in either or both of the organisms being compared. Chervitz *et al.* (*120*) deal with this case by analyzing a phylogenetic tree that described the relationships between all of the sequences in both organisms, and then looked for pairs of genes that were nearest neighbors in the tree. If the nearest-neighbor pairs were from different organisms, those genes were presumed to be orthologs. We note that these nearest neighbors can often be confidently identified from pairwise sequence comparison without having to examine a phylogenetic tree (see legend to Fig. 16). If the nearest neighbors are not from different organisms, there has been a paralogous expansion in one or both organisms after the speciation event (and/or a gene loss by one organism). When this one-to-one correspondence is lost, defining an ortholog becomes ambiguous. For our initial computational overview of the predicted human protein set, we could not answer this question for every predicted protein. Therefore, we consider only "strict orthologs," i.e., the proteins with unambiguous one-to-one relationships (Fig. 16). By these criteria, there are 2758 strict human-fly orthologs, 2031 human-worm (1523 in common between these sets). We define the evolutionarily conserved set as those 1523 human proteins that have strict orthologs in both *D. melanogaster* and *C. elegans*.

The distribution of the functions of the conserved protein set is shown in Fig. 16. Comparison with Fig. 15 shows that, not surprisingly, the set of conserved proteins is not distributed among molecular functions in the same way as the whole human protein set. Compared with the whole human set (Fig. 15), there are several categories that are overrepresented in the conserved set by a factor of ~2 or more. The first category is nucleic acid enzymes, primarily the transcriptional machinery (notably DNA/RNA methyltransferases, DNA/RNA polymerases, helicases, DNA ligases, DNA- and RNA-processing factors, nucleases, and ribosomal proteins). The basic transcriptional and translational machinery is well known to have been conserved over evolution, from bacteria through to the most complex eukaryotes. Many ribonucleoproteins involved in RNA splicing also appear to be conserved among the animals. Other enzyme types are also overrepresented (transferases, oxidoreductases, ligases, lyases, and isomerases). Many of these en-

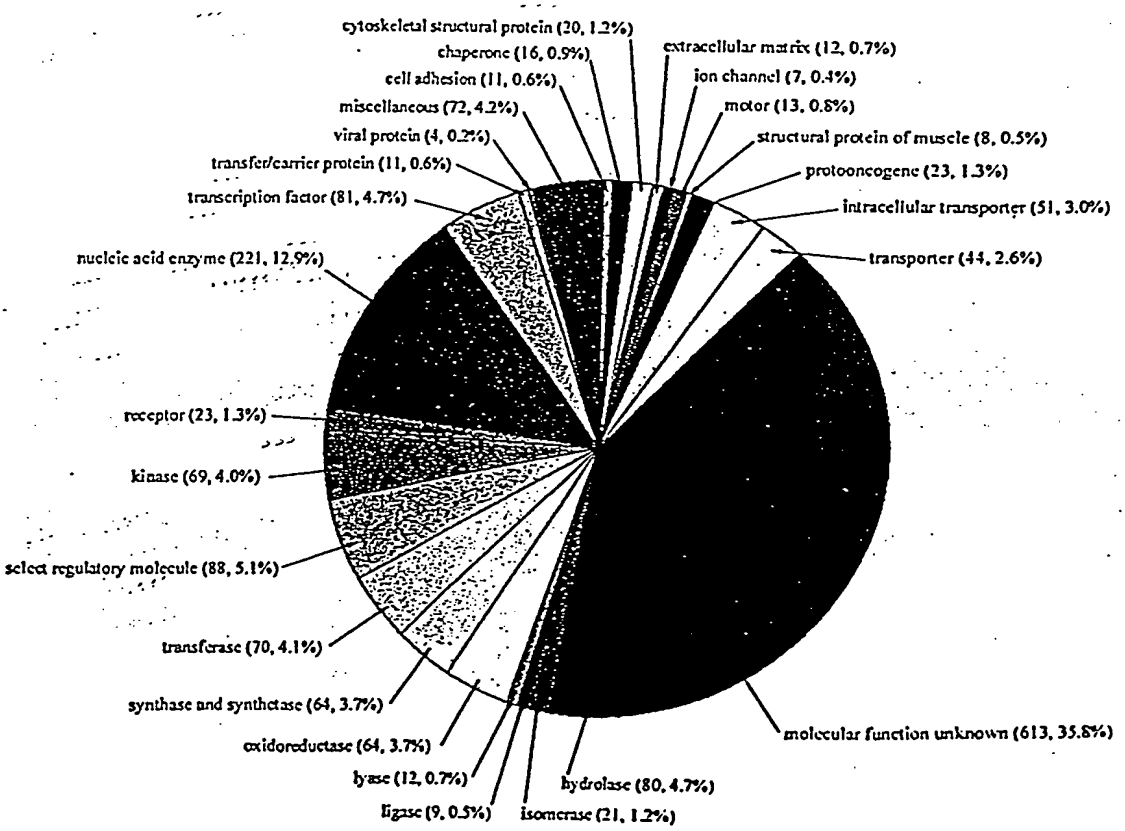


Fig. 16. Functions of putative orthologs across vertebrate and invertebrate genomes. Each slice lists the number and percentages (in parentheses) of "strict orthologs" between the human, fly, and worm genomes involved in a given category of molecular function. "Strict orthologs" are defined here as bi-directional BLAST best hits (*180*) such that each orthologous pair (i) has a BLASTP *P*-value of $\leq 10^{-10}$ (*120*), and (ii) has a more significant BLASTP score than any paralogs in either organism, i.e., there has likely been no duplication subsequent to speciation that might make the orthology ambiguous. This measure is quite strict and is a lower bound on the number of orthologs. By these criteria, there are 2758 strict human-fly orthologs, and 2031 human-worm orthologs (1523 in common between these sets).

zymes are involved in intermediary metabolism. The only exception is the hydrolase category, which is not significantly overrepresented in the shared protein set. Proteases form the largest part of this category, and several large protease families have expanded in each of these three organisms after their divergence. The category of select regulatory molecules is also overrepresented in the conserved set. The major conserved families are small guanosine triphosphatases (GTPases) (especially the Ras-related superfamily, including ADP ribosylation factor) and cell cycle regulators (particularly the cullin family, cyclin C family, and several cell division protein kinases). The last two significantly overrepresented categories are protein transport and trafficking, and chaperones. The most conserved groups in these categories are proteins involved in coated vesicle-mediated transport, and chaperones involved in protein folding and heat-shock response [particularly the DNAJ family, and heat-shock protein 60 (HSP60), HSP70, and HSP90 families]. These observations provide only a conservative estimate of the protein families in the context of specific cellular processes that were likely derived from the last common ancestor of the human, fly, and worm. As stated before, this analysis does not provide a complete estimate of conservation across the three animal genomes, as paralogous duplication makes the determination of true orthologs difficult within the members of conserved protein families.

7.3 Differences between the human genome and other sequenced eukaryotic genomes

To explore the molecular building blocks of the vertebrate taxon, we have compared the human genome with the other sequenced eukaryotic genomes at three levels: molecular functions, protein families, and protein domains.

Molecular differences can be correlated with phenotypic differences to begin to reveal the developmental and cellular processes that are unique to the vertebrates. Tables 18 and 19 display a comparison among all sequenced eukaryotic genomes, over selected protein/domain families (defined by sequence similarity, e.g., the serine-threonine protein kinases) and superfamilies (defined by shared molecular function, which may include several sequence-related families, e.g., the cytokines). In these tables we have focused on (super) families that are either very large or that differ significantly in humans compared with the other sequenced eukaryote genomes. We have found that the most prominent human expansions are in proteins involved in (i) acquired immune functions; (ii) neural development, structure, and functions; (iii) intercellular and intracellular signaling pathways in development and homeostasis; (iv) hemostasis; and (v) apoptosis.

Acquired immunity. One of the most striking differences between the human genome and the *Drosophila* or *C. elegans* genome is the appearance of genes involved in acquired immunity (Tables 18 and 19). This is expected, because the acquired immune response is a defense system that only occurs in vertebrates. We observe 22 class I and 22 class II major histocompatibility complex (MHC) antigen genes and 114 other immunoglobulin genes in the human genome. In addition, there are 59 genes in the cognate immunoglobulin receptor family. At the domain level, this is exemplified by an expansion and recruitment of the ancient immunoglobulin fold to constitute molecules such as MHC, and of the integrin fold to form several of the cell adhesion molecules that mediate interactions between immune effector cells and the extracellular matrix. Vertebrate-specific proteins include the paracrine immune regulators family of secreted 4-alpha helical bundle proteins, namely the cytokines and chemokines. Some of the cytoplasmic signal transduction components associated with cytokine receptor signal transduction are also features that are poorly represented in the fly and worm. These include protein domains found in the signal transducer and activator of transcription (STATs), the suppressors of cytokine signaling (SOCS), and protein inhibitors of activated STATs (PIAS). In contrast, many of the animal-specific protein domains that play a role in innate immune response, such as the Toll receptors, do not appear to be significantly expanded in the human genome.

Neural development, structure, and function. In the human genome, as compared with the worm and fly genomes, there is a marked increase in the number of members of protein families that are involved in neural development. Examples include neurotrophic factors such as ependymin, nerve growth factor, and signaling molecules such as semaphorins, as well as the number of proteins involved directly in neural structure and function such as myelin proteins, voltage-gated ion channels, and synaptic proteins such as synaptotagmin. These observations correlate well with the known phenotypic differences between the nervous systems of these taxa, notably (i) the increase in the number and connectivity of neurons; (ii) the increase in number of distinct neural cell types (as many as a thousand or more in human compared with a few hundred in fly and worm) (*121*); (iii) the increased length of individual axons; and (iv) the significant increase in glial cell number, especially the appearance of myelinating glial cells, which are electrically inert supporting cells differentiated from the same stem cells as neurons. A number of prominent protein expansions are involved in the processes of neural development. Of the extracellular domains that mediate cell adhesion, the connexin domain–containing proteins (*122*) exist only in humans. These proteins, which are not present in the *Drosophila* or *C. elegans* genomes, appear to provide the constitutive subunits of intercellular channels and the structural basis for electrical coupling. Pathway finding by axons and neuronal network formation is mediated through a subset of ephrins and their cognate receptor tyrosine kinases that act as positional labels to establish topographical projections (*123*). The probable biological role for the semaphorins (22 in human compared with 6 in the fly and 2 in the worm) and their receptors (neuropilins and plexins) is that of axonal guidance molecules (*124*). Signaling molecules such as neurotrophic factors and some cytokines have been shown to regulate neuronal cell survival, proliferation, and axon guidance (*125*). Notch receptors and ligands play important roles in glial cell fate determination and gliogenesis (*126*).

Other human expanded gene families play key roles directly in neural structure and function. One example is synaptotagmin (expanded more than twofold in humans relative to the invertebrates), originally found to regulate synaptic transmission by serving as a Ca^{2+} sensor (or receptor) during synaptic vesicle fusion and release (*127*). Of interest is the increased co-occurrence in humans of PDZ and the SH3 domains in neuronal-specific adaptor molecules; examples include proteins that likely modulate channel activity at synaptic junctions (*128*). We also noted expansions in several ion-channel families (Table 19), including the EAG subfamily (related to cyclic nucleotide gated channels), the voltage-gated calcium/sodium channel family, the inward-rectifier potassium channel family, and the voltage-gated potassium channel, alpha subunit family. Voltage-gated sodium and potassium channels are involved in the generation of action potentials in neurons. Together with voltage-gated calcium channels, they also play a key role in coupling action potentials to neurotransmitter release, in the development of neurites, and in short-term memory. The recent observation of a calcium-regulated association between sodium channels and synaptotagmin may have consequences for the establishment and regulation of neuronal excitability (*129*).

Myelin basic protein and myelin-associated glycoprotein are major classes of protein components in both the central and peripheral nervous system of vertebrates. Myelin P0 is a major component of peripheral myelin, and myelin proteolipid and myelin oligodendrocyte glycopotein are found in the central nervous system. Mutations in any of these

Table 18. Domain-based comparative analysis of proteins in *H. sapiens* (H), *D. melanogaster* (F), *C. elegans* (W), *S. cerevisiae* (Y), and *A. thaliana* (A). The predicted protein set of each of the above eukaryotic organisms was analyzed with Pfam version 5.5 using E-value cutoffs of 0.001. The number of proteins containing the specified Pfam domains as well as the total number of domains (in parentheses) are shown in each column. Domains were categorized into cellular processes for presentation. Some domains (i.e., SH2) are listed in more than one cellular process. Results of the Pfam analysis may differ from results obtained based on human curation of protein families, owing to the limitations of large-scale automatic classifications. Representative examples of domains with reduced counts owing to the stringent E value cutoff used for this analysis are marked with a double asterisk (**). Examples include short divergent and predominantly alpha-helical domains, and certain classes of cysteine-rich zinc finger proteins.

Accession number	Domain name	Domain description	H	F	W	Y	A
		Developmental and homeostatic regulators					
PF02039	Adrenomedullin	Adrenomedullin	1	0	0	0	0
PF00212	ANP	Atrial natriuretic peptide	2	0	0	0	0
PF00028	Cadherin	Cadherin domain	100 (550)	14 (157)	16 (66)	0	0
PF00214	Calc_CGRP_IAPP	Calcitonin/CGRP/IAPP family	3	0	0	0	0
PF01110	CNTF	Ciliary neurotrophic factor	1	0	0	0	0
PF01093	Clusterin	Clusterin	3	0	0	0	0
PF00029	Connexin	Connexin	14 (16)	0	0	0	0
PF00976	ACTH_domain	Corticotropin ACTH domain	1	0	0	0	0
PF00473	CRF	Corticotropin-releasing factor family	2	1	0	0	0
PF00007	Cys_knot	Cystine-knot domain	10 (11)	2	0	0	0
PF00778	DIX	Dix domain	5	2	4	0	0
PF00322	Endothelin	Endothelin family	3	0	0	0	0
PF00812	Ephrin	Ephrin	7 (8)	2	4	0	0
PF01404	EPh_lbd	Ephrin receptor ligand binding domain	12	2	1	0	0
PF00167	FGF	Fibroblast growth factor	23	1	1	0	0
PF01534	Frizzled	Frizzled/Smoothened family membrane region	9	7	3	0	0
PF00236	Hormone6	Glycoprotein hormones	1	0	0	0	0
PF01153	Glypican	Glypican	14	2	1	0	0
PF01271	Granin	Grainin (chromogranin or secretogranin)	3	0	0	0	0
PF02058	Guanylin	Guanylin precursor	1	0	0	0	0
PF00049	Insulin	Insulin/IGF/Relaxin family	7	4	0	0	0
PF00219	IGFBP	Insulin-like growth factor binding proteins	10	0	0	0	0
PF02024	Leptin	Leptin	1	0	0	0	0
PF00193	Xlink	LINK (hyaluron binding)	13 (23)	0	1	0	0
PF00243	NGF	Nerve growth factor family	3	0	0	0	0
PF02158	Neuregulin	Neuregulin family	4	0	0	0	0
PF00184	Hormone5	Neurohypophysial hormones	1	0	0	0	0
PF02070	NMU	Neuromedin U	1	0	0	0	0
PF00066	Notch	Notch (DSL) domain	3 (5)	2 (4)	2 (6)	0	0
PF00865	Osteopontin	Osteopontin	1	0	0	0	0
PF00159	Hormone3	Pancreatic hormone peptides	3	0	0	0	0
PF01279	Parathyroid	Parathyroid hormone family	2	0	0	0	0
PF00123	Hormone2	Peptide hormone	5 (9)	0	0	0	0
PF00341	PDGF	Platelet-derived growth factor (PDGF)	5	1	0	0	0
PF01403	Sema	Sema domain	27 (29)	8 (10)	3 (4)	0	0
PF01033	Somatomedin_B	Somatomedin B domain	5 (8)	3	0	0	0
PF00103	Hormone	Somatotropin	1	0	0	0	0
PF02208	Sorb	Sorbin homologous domain	2	0	0	0	0
PF02404	SCF	Stem cell factor	2	0	0	0	0
PF01034	Syndecan	Syndecan domain	3	1	1	0	0
PF00020	TNFR_c6	TNFR/NGFR cysteine-rich region	17 (31)	1	0	0	0
PF00019	TGF-β	Transforming growth factor β-like domain	27 (28)	6	4	0	0
PF01099	Uteroglobin	Uteroglobin family	3	0	0	0	0
PF01160	Opioids_neuropep	Vertebrate endogenous opioids neuropeptide	3	0	0	0	0
PF00110	Wnt	Wnt family of developmental signaling proteins	18	7 (10)	5	0	0
		Hemostasis					
PF01821	ANATO	Anaphylotoxin-like domain	6 (14)	0	0	0	0
PF00386	C1q	C1q domain	24	0	0	0	0
PF00200	Disintegrin	Disintegrin	18	2	3	0	0
PF00754	F5_F8_type_C	F5/8 type C domain	15 (20)	5 (6)	2	0	0
PF01410	COLFI	Fibrillar collagen C-terminal domain	10	0	0	0	0
PF00039	Fn1	Fibronectin type I domain	5 (18)	0	0	0	0
PF00040	Fn2	Fibronectin type II domain	11 (16)	0	0	0	0
PF00051	Kringle	Kringle domain	15 (24)	2	2	0	0
PF01823	MACPF	MAC/Perforin domain	6	0	0	0	0
PF00354	Pentaxin	Pentaxin family	9	0	0	0	0
PF00277	SAA_proteins	Serum amyloid A protein	4	0	0	0	0
PF00084	Sushi	Sushi domain (SCR repeat)	53 (191)	11 (42)	8 (45)	0	0
PF02210	TSPN	Thrombospondin N-terminal–like domains	14	1	0	0	0
PF01108	Tissue_fac	Tissue factor	1	0	0	0	0
PF00868	Transglutamin_N	Transglutaminase family	6	1	0	0	0
PF00927	Transglutamin_C	Transglutaminase family	8	1	0	0	0

Table 18 (*Continued*)

Accession number	Domain name	Domain description	H	F	W	Y	A
PF00594	Gla	Vitamin K-dependent carboxylation/gamma-carboxyglutamic (GLA) domain	11	0	0	0	0
		Immune response					
PF00711	Defensin_beta	Beta defensin	1	0	0	0	0
PF00748	Calpain_inhib	Calpain inhibitor repeat	3 (9)	0	0	0	0
PF00666	Cathelicidins	Cathelicidins	2	0	0	0	0
PF00129	MHC_I	Class I histocompatibility antigen, domains alpha 1 and 2	18 (20)	0	0	0	0
PF00993	MHC_II_alpha**	Class II histocompatibility antigen, alpha domain	5 (6)	0	0	0	0
PF00969	MHC_II_beta**	Class II histocompatibility antigen, beta domain	7	0	0	0	0
PF00879	Defensin_propep	Defensin propeptide	3	0	0	0	0
PF01109	GM_CSF	Granulocyte-macrophage colony-stimulating factor	1	0	0	0	0
PF00047	Ig	Immunoglobulin domain	381 (930)	125 (291)	67 (323)	0	0
PF00143	Interferon	Interferon alpha/beta domain	7 (9)	0	0	0	0
PF00714	IFN-gamma	Interferon gamma	1	0	0	0	0
PF00726	IL10	Interleukin-10	1	0	0	0	0
PF02372	IL15	Interleukin-15	1	0	0	0	0
PF00715	IL2	Interleukin-2	1	0	0	0	0
PF00727	IL4	Interleukin-4	1	0	0	0	0
PF02025	IL5	Interleukin-5	1	0	0	0	0
PF01415	IL7	Interleukin-7/9 family	1	0	0	0	0
PF00340	IL1	Interleukin-1	1	0	0	0	0
PF02394	IL1_propep	Interleukin-1 propeptide	7	0	0	0	0
PF02059	IL3	Interleukin-3	1	0	0	0	0
PF00489	IL6	Interleukin-6/G-CSF/MGF family	1	0	0	0	0
PF01291	LIF_OSM	Leukemia inhibitory factor (LIF)/oncostatin (OSM) family	2	0	0	0	0
PF00323	Defensins	Mammalian defensin	2	0	0	0	0
PF01091	PTN_MK	PTN/MK heparin-binding protein	2	0	0	0	0
PF00277	SAA_proteins	Serum amyloid A protein	4	0	0	0	0
PF00048	IL8	Small cytokines (intecrine/chemokine), interleukin-8 like	32	0	0	0	0
PF01582	TIR	TIR domain	18	8	2	0	131 (143)
PF00229	TNF	TNF (tumor necrosis factor) family	12	0	0	0	0
PF00088	Trefoil	Trefoil (P-type) domain	5 (6)	0	2	0	0
		PI-PY-rho GTPase signaling					
PF00779	BTK	BTK motif	5	1	0	0	0
PF00168	C2	C2 domain	73 (101)	32 (44)	24 (35)	6 (9)	66 (90)
PF00609	DAGKa	Diacylglycerol kinase accessory domain (presumed)	9	4	7	0	6
PF00781	DAGKc	Diacylglycerol kinase catalytic domain (presumed)	10	8	8	2	11 (12)
PF00610	DEP	Domain found in Dishevelled, Egl-10, and Pleckstrin (DEP)	12 (13)	4	10	5	2
PF01363	FYVE	FYVE zinc finger	28 (30)	14	15	5	15
PF00996	GDI	GDP dissociation inhibitor	6	2	1	1	3
PF00503	G-alpha	G-protein alpha subunit	27 (30)	10	20 (23)	2	5
PF00631	G-gamma	G-protein gamma like domains	16	5	5	1	0
PF00616	RasGAP	GTPase-activator protein for Ras-like GTPase	11	5	8	3	0
PF00618	RasGEFN	Guanine nucleotide exchange factor for Ras-like GTPases; N-terminal motif	9	2	3	5	0
PF00625	Guanylate_kin	Guanylate kinase	12	8	7	1	4
PF02189	ITAM	Immunoreceptor tyrosine-based activation motif	3	0	0	0	0
PF00169	PH	PH domain	193 (212)	72 (78)	65 (68)	24	23
PF00130	DAG_PE-bind	Phorbol esters/diacylglycerol binding domain (C1 domain)	45 (56)	25 (31)	26 (40)	1 (2)	4
PF00388	PI-PLC-X	Phosphatidylinositol-specific phospholipase C, X domain	12	3	7	1	8
PF00387	PI-PLC-Y	Phosphatidylinositol-specific phospholipase C, Y domain	11	2	7	1	8
PF00640	PID	Phosphotyrosine Interaction domain (PTB/PID)	24 (27)	13	11 (12)	0	0
PF02192	PI3K_p85B	PI3-kinase family, p85-binding domain	2	1	1	0	0
PF00794	PI3K_rbd	PI3-kinase family, ras-binding domain	6	3	1	0	0
PF01412	ArfGAP	Putative GTP-ase activating protein for Arf	16	9	8	6	15
PF02196	RBD	Raf-like Ras-binding domain	6 (7)	4	1	0	0
PF02145	Rap_GAP	Rap/ran-GAP	5	4	2	0	0
PF00788	RA	Ras association (RalGDS/AF-6) domain	18 (19)	7 (9)	6	1	0
PF00071	Ras	Ras family	126	56 (57)	51	23	78
PF00617	RasGEF	RasGEF domain	21	8	7	5	0
PF00615	RGS	Regulator of G protein signaling domain	27	6 (7)	12 (13)	1	0
PF02197	RIIa	Regulatory subunit of type II PKA R-subunit	4	1	2	1	0

Table 18 (*Continued*)

Accession number	Domain name	Domain description	H	F	W	Y	A
PF00620	RhoGAP	RhoGAP domain	59	19	20	9	8
PF00621	RhoGEF	RhoGEF domain	46	23 (24)	18 (19)	3	0
PF00536	SAM	SAM domain (Sterile alpha motif)	29 (31)	15	8	3	6
PF01369	Sec7	Sec7 domain	13	5	5	5	9
PF00017	SH2	Src homology 2 (SH2) domain	87 (95)	33 (39)	44 (48)	1	3
PF00018	SH3	Src homology 3 (SH3) domain	143 (182)	55 (75)	46 (61)	23 (27)	4
PF01017	STAT	STAT protein	7	1	1 (2)	0	0
PF00790	VHS	VHS domain	4	2	4	4	8
PF00568	WH1	WH1 domain	7	2	2 (3)	1	0
		Domains involved in apoptosis					
PF00452	Bcl-2	Bcl-2	9	2	1	0	0
PF02180	BH4	Bcl-2 homology region 4	3	0	1	0	0
PF00619	CARD	Caspase recruitment domain	16	0	2	0	0
PF00531	Death	Death domain	16	5	7	0	0
PF01335	DED	Death effector domain	4 (5)	0	0	0	0
PF02179	BAG	Domain present in Hsp70 regulators	5 (8)	3	2	1	5
PF00656	ICE_p20	ICE-like protease (caspase) p20 domain	11	7	3	0	0
PF00653	BIR	Inhibitor of Apoptosis domain	8 (14)	5 (9)	2 (3)	1 (2)	0
		Cytoskeletal					
PF00022	Actin	Actin	61 (64)	15 (16)	12	9 (11)	24
PF00191	Annexin	Annexin	16 (55)	4 (16)	4 (11)	0	6 (16)
PF00402	Calponin	Calponin family	13 (22)	3	7 (19)	0	0
PF00373	Band_41	FERM domain (Band 4.1 family)	29 (30)	17 (19)	11 (14)	0	0
PF00880	Nebulin_repeat	Nebulin repeat	4 (148)	1 (2)	1	0	0
PF00681	Plectin_repeat	Plectin repeat	2 (11)	0	0	0	0
PF00435	Spectrin	Spectrin repeat	31 (195)	13 (171)	10 (93)	0	0
PF00418	Tubulin-binding	Tau and MAP proteins, tubulin-binding	4 (12)	1 (4)	2 (8)	0	0
PF00992	Troponin	Troponin	4	6	8	0	0
PF02209	VHP	Villin headpiece domain	5	2	2	0	5
PF01044	Vinculin	Vinculin family	4	2	1	0	0
		ECM adhesion					
PF01391	Collagen	Collagen triple helix repeat (20 copies)	65 (279)	10 (46)	174 (384)	0	0
PF01413	C4	C-terminal tandem repeated domain in type 4 procollagen	6 (11)	2 (4)	3 (6)	0	0
PF00431	CUB	CUB domain	47 (69)	9 (47)	43 (67)	0	0
PF00008	EGF	EGF-like domain	108 (420)	45 (186)	54 (157)	0	1
PF00147	Fibrinogen_C	Fibrinogen beta and gamma chains, C-terminal globular domain	26	10 (11)	6	0	0
PF00041	Fn3	Fibronectin type III domain	106 (545)	42 (168)	34 (156)	0	1
PF00757	Furin-like	Furin-like cysteine rich region	5	2	1	0	0
PF00357	Integrin_A	Integrin alpha cytoplasmic region	3	1	2	0	0
PF00362	Integrin_B	Integrins, beta chain	8	2	2	0	0
PF00052	Laminin_B	Laminin B (Domain IV)	8 (12)	4 (7)	6 (10)	0	0
PF00053	Laminin_EGF	Laminin EGF-like (Domains III and V)	24 (126)	9 (62)	11 (65)	0	0
PF00054	Laminin_G	Laminin G domain	30 (57)	18 (42)	14 (26)	0	0
PF00055	Laminin_Nterm	Laminin N-terminal (Domain VI)	10	6	4	0	0
PF00059	Lectin_c	Lectin C-type domain	47 (76)	23 (24)	91 (132)	0	0
PF01463	LRRCT	Leucine rich repeat C-terminal domain	69 (81)	23 (30)	7 (9)	0	0
PF01462	LRRNT	Leucine rich repeat N-terminal domain	40 (44)	7 (13)	3 (6)	0	0
PF00057	Ldl_recept_a	Low-density lipoprotein receptor domain class A	35 (127)	33 (152)	27 (113)	0	0
PF00058	Ldl_recept_b	Low-density lipoprotein receptor repeat class B	15 (96)	9 (56)	7 (22)	0	0
PF00530	SRCR	Scavenger receptor cysteine-rich domain	11 (46)	4 (8)	1 (2)	0	0
PF00084	Sushi	Sushi domain (SCR repeat)	53 (191)	11 (42)	8 (45)	0	0
PF00090	Tsp_1	Thrombospondin type 1 domain	41 (66)	11 (23)	18 (47)	0	0
PF00092	Vwa	von Willebrand factor type A domain	34 (58)	0	17 (19)	0	1
PF00093	Vwc	von Willebrand factor type C domain	19 (28)	6 (11)	2 (5)	0	0
PF00094	Vwd	von Willebrand factor type D domain	15 (35)	3 (7)	9	0	0
		Protein interaction domains					
PF00244	14-3-3	14-3-3 proteins	20	3	3	2	15
PF00023	Ank	Ank repeat	145 (404)	72 (269)	75 (223)	12 (20)	66 (111)
PF00514	Armadillo_seg	Armadillo/beta-catenin-like repeats	22 (56)	11 (38)	3 (11)	2 (10)	25 (67)
PF00168	C2	C2 domain	73 (101)	32 (44)	24 (35)	6 (9)	66 (90)
PF00027	cNMP_binding	Cyclic nucleotide-binding domain	26 (31)	21 (33)	15 (20)	2 (3)	22
PF01556	DnaJ_C	DnaJ C terminal region	12	9	5	3	19
PF00226	DnaJ	DnaJ domain	44	34	33	20	93
PF00036	Efhand**	EF hand	83 (151)	64 (117)	41 (86)	4 (11)	120 (328)
PF00611	FCH	Fes/CIP4 homology domain	9	3	2	4	0
PF01846	FF	FF domain	4 (11)	4 (10)	3 (16)	2 (5)	4 (8)
PF00498	FHA	FHA domain	13	15	7	13 (14)	17

myelin proteins result in severe demyelination, which is a pathological condition in which the myelin is lost and the nerve conduction is severely impaired (*130*). Humans have at least 10 genes belonging to four different families involved in myelin production (five myelin P0, three myelin proteolipid, myelin basic protein, and myelin-oligodendrocyte glycoprotein, or MOG), and possibly more-remotely related members of the MOG family. Flies have only a single myelin proteolipid, and worms have none at all.

Intercellular and intracellular signaling pathways in development and homeostasis. Many protein families that have expanded in humans relative to the invertebrates are involved in signaling processes, particularly in response to development and differentiation

Table 18 (*Continued*)

Accession number	Domain name	Domain description	H	F	W	Y	A
PF00254	FKBP	FKBP-type peptidyl-prolyl cis-trans Isomerases	15 (20)	7 (8)	7 (13)	4	24 (29)
PF01590	GAF	GAF domain	7 (8)	2 (4)	1	0	10
PF01344	Kelch	Kelch motif	54 (157)	12 (48)	13 (41)	3	102 (178)
PF00560	LRR**	Leucine Rich Repeat	25 (30)	24 (30)	7 (11)	1	15 (16)
PF00917	MATH	MATH domain	11	5	88 (161)	1	61 (74)
PF00989	PAS	PAS domain	18 (19)	9 (10)	6	1	13 (18)
PF00595	PDZ	PDZ domain (Also known as DHR or GLGF)	96 (154)	60 (87)	46 (66)	2	5
PF00169	PH	PH domain	193 (212)	72 (78)	65 (68)	24	23
PF01535	PPR**	PPR repeat	5	3 (4)	0	1	474 (2485)
PF00536	SAM	SAM domain (Sterile alpha motif)	29 (31)	15	8	3	6
PF01369	Sec7	Sec7 domain	13	5	5	5	9
PF00017	SH2	Src homology 2 (SH2) domain	87 (95)	33 (39)	44 (48)	1	3
PF00018	SH3	Src homology 3 (SH3) domain	143 (182)	55 (75)	46 (61)	23 (27)	4
PF01740	STAS	STAS domain	5	1	6	2	13
PF00515	TPR**	TPR domain	72 (131)	39 (101)	28 (54)	16 (31)	65 (124)
PF00400	WD40**	WD40 domain	136 (305)	98 (226)	72 (153)	56 (121)	167 (344)
PF00397	WW	WW domain	32 (53)	24 (39)	16 (24)	5 (8)	11 (15)
PF00569	ZZ	ZZ-Zinc finger present in dystrophin, CBP/p300	10 (11)	13	10	2	10
		Nuclear Interaction domains					
PF01754	Zf-A20	A20-like zinc finger	2 (8)	2	2	0	8
PF01388	ARID	ARID DNA binding domain	11	6	4	2	7
PF01426	BAH	BAH domain	8 (10)	7 (8)	4 (5)	5	21 (25)
PF00643	Zf-B_box**	B-box zinc finger	32 (35)	1	2	0	0
PF00533	BRCT	BRCA1 C Terminus (BRCT) domain	17 (28)	10 (18)	23 (35)	10 (16)	12 (16)
PF00439	Bromodomain	Bromodomain	37 (48)	16 (22)	18 (26)	10 (15)	28
PF00651	BTB	BTB/POZ domain	97 (98)	62 (64)	86 (91)	1 (2)	30 (31)
PF00145	DNA_methylase	C-5 cytosine-specific DNA methylase	3 (4)	1	0	0	13 (15)
PF00385	Chromo	chromo' (CHRomatin Organization MOdifier) domain	24 (27)	14 (15)	17 (18)	1 (2)	12
PF00125	Histone	Core histone H2A/H2B/H3/H4	75 (81)	5	71 (73)	8	48
PF00134	Cyclin	Cyclin	19	10	10	11	35
PF00270	DEAD	DEAD/DEAH box helicase	63 (66)	48 (50)	55 (57)	50 (52)	84 (87)
PF01529	Zf-DHHC	DHHC zinc finger domain	15	20	16	7	22
PF00646	F-box**	F-box domain	16	15	309 (324)	9	165 (167)
PF00250	Fork_head	Fork head domain	35 (36)	20 (21)	15	4	0
PF00320	GATA	GATA zinc finger	11 (17)	5(6)	8 (10)	9	26
PF01585	G-patch	G-patch domain	18	16	13	4	14 (15)
PF00010	HLH**	Helix-loop-helix DNA-binding domain	60 (61)	44	24	4	39
PF00850	Hist_deacetyl	Histone deacetylase family	12	5 (6)	8 (10)	5	10
PF00046	Homeobox	Homeobox domain	160 (178)	100 (103)	82 (84)	6	66
PF01833	TIG	IPT/TIG domain	29 (53)	11 (13)	5 (7)	2	1
PF02373	JmjC	JmjC domain	10	4	6	4	7
PF02375	JmjN	JmjN domain	7	4	2	3	7
PF00013	KH-domain	KH domain	28 (67)	14 (32)	17 (46)	4 (14)	27 (61)
PF01352	KRAB	KRAB box	204 (243)	0	0	0	0
PF00104	Hormone_rec	Ligand-binding domain of nuclear hormone receptor	47	17	142 (147)	0	0
PF00412	LIM	LIM domain containing proteins	62 (129)	33 (83)	33 (79)	4 (7)	10 (16)
PF00917	MATH	MATH domain	11	5	88 (161)	1	61 (74)
PF00249	Myb_DNA-binding	Myb-like DNA-binding domain	32 (43)	18 (24)	17 (24)	15 (20)	243 (401)
PF02344	Myc-LZ	Myc leucine zipper domain	1	0	0	0	0
PF01753	Zf-MYND	MYND finger	14	14	9	1	7
PF00628	PHD	PHD-finger	68 (86)	40 (53)	32 (44)	14 (15)	96 (105)
PF00157	Pou	Pou domain—N-terminal to homeobox domain	15	5	4	0	0
PF02257	RFX_DNA_binding	RFX DNA-binding domain	7	2	1	1	0
PF00076	Rrm	RNA recognition motif (a.k.a. RRM, RBD, or RNP domain)	224 (324)	127 (199)	94 (145)	43 (73)	232 (369)
PF02037	SAP	SAP domain	15	8	5	5	6 (7)
PF00622	SPRY	SPRY domain	44 (51)	10 (12)	5 (7)	3	6
PF01852	START	START domain	10	2	6	0	23
PF00907	T-box	T-box	17 (19)	8	22	0	0

Table 18 (*Continued*)

Accession number	Domain name	Domain description	H	F	W	Y	A
PF02135	Zf-TAZ	TAZ finger	2 (3)	1 (2)	6 (7)	0	10 (15)
PF01285	TEA	TEA domain	4	1	1	1	0
PF02176	Zf-TRAF	TRAF-type zinc finger	6 (9)	1 (3)	1	0	2
PF00352	TBP	Transcription factor TFIID (or TATA-binding protein, TBP)	2 (4)	4 (8)	2 (4)	1 (2)	2 (4)
PF00567	TUDOR	TUDOR domain	9 (24)	9 (19)	4 (5)	0	2
PF00642	Zf-CCCH	Zinc finger C-x8-C-x5-C-x3-H type (and similar)	17 (22)	6 (8)	22 (42)	3 (5)	31 (46)
PF00096	Zf-C2H2**	Zinc finger, C2H2 type	564 (4500)	234 (771)	68 (155)	34 (56)	21 (24)
PF00097	Zf-C3HC4	Zinc finger, C3HC4 type (RING finger)	135 (137)	57	88 (89)	18	298 (304)
PF00098	Zf-CCHC	Zinc knuckle	9 (17)	6 (10)	17 (33)	7 (13)	68 (91)

(Tables 18 and 19). They include secreted hormones and growth factors, receptors, intracellular signaling molecules, and transcription factors.

Developmental signaling molecules that are enriched in the human genome include growth factors such as wnt, transforming growth factor–β (TGF-β), fibroblast growth factor (FGF), nerve growth factor, platelet derived growth factor (PDGF), and ephrins. These growth factors affect tissue differentiation and a wide range of cellular processes involving actin-cytoskeletal and nuclear regulation. The corresponding receptors of these developmental ligands are also expanded in humans. For example, our analysis suggests at least 8 human ephrin genes (2 in the fly, 4 in the worm) and 12 ephrin receptors (2 in the fly, 1 in the worm). In the wnt signaling pathway, we find 18 wnt family genes (6 in the fly, 5 in the worm) and 12 frizzled receptors (6 in the fly, 5 in the worm). The Groucho family of transcriptional corepressors downstream in the wnt pathway are even more markedly expanded, with 13 predicted members in humans (2 in the fly, 1 in the worm).

Extracellular adhesion molecules involved in signaling are expanded in the human genome (Tables 18 and 19). The interactions of several of these adhesion domains with extracellular matrix proteoglycans play a critical role in host defense, morphogenesis, and tissue repair (*131*). Consistent with the well-defined role of heparan sulfate proteoglycans in modulating these interactions (*132*), we observe an expansion of the heparin sulfate sulfotransferases in the human genome relative to worm and fly. These sulfotransferases modulate tissue differentiation (*133*). A similar expansion in humans is noted in structural proteins that constitute the actin-cytoskeletal architecture. Compared with the fly and worm, we observe an explosive expansion of the nebulin (35 domains per protein on average), aggrecan (12 domains per protein on average), and plectin (5 domains per protein on average) repeats in humans. These repeats are present in proteins involved in modulating the actin-cytoskeleton with predominant expression in neuronal, muscle, and vascular tissues.

Comparison across the five sequenced eukaryotic organisms revealed several expanded protein families and domains involved in cytoplasmic signal transduction (Table 18). In particular, signal transduction pathways playing roles in developmental regulation and acquired immunity were substantially enriched. There is a factor of 2 or greater expansion in humans in the Ras superfamily GTPases and the GTPase activator and GTP exchange factors associated with them. Although there are about the same number of tyrosine kinases in the human and *C. elegans* genomes, in humans there is an increase in the SH2, PTB, and ITAM domains involved in phosphotyrosine signal transduction. Further, there is a twofold expansion of phosphodiesterases in the human genome compared with either the worm or fly genomes.

The downstream effectors of the intracellular signaling molecules include the transcription factors that transduce developmental fates. Significant expansions are noted in the ligand-binding nuclear hormone receptor class of transcription factors compared with the fly genome, although not to the extent observed in the worm (Tables 18 and 19). Perhaps the most striking expansion in humans is in the C2H2 zinc finger transcription factors. Pfam detects a total of 4500 C2H2 zinc finger domains in 564 human proteins, compared with 771 in 234 fly proteins. This means that there has been a dramatic expansion not only in the number of C2H2 transcription factors, but also in the number of these DNA-binding motifs per transcription factor (8 on average in humans, 3.3 on average in the fly, and 2.3 on average in the worm). Furthermore, many of these transcription factors contain either the KRAB or SCAN domains, which are not found in the fly or worm genomes. These domains are involved in the oligomerization of transcription factors and increase the combinatorial partnering of these factors. In general, most of the transcription factor domains are shared between the three animal genomes, but the reassortment of these domains results in organism-specific transcription factor families. The domain combinations found in the human, fly, and worm include the BTB with C2H2 in the fly and humans, and homeodomains alone or in combination with Pou and LIM domains in all of the animal genomes. In plants, however, a different set of transcription factors are expanded, namely, the myb family, and a unique set that includes VP1 and AP2 domain–containing proteins (*134*). The yeast genome has a paucity of transcription factors compared with the multicellular eukaryotes, and its repertoire is limited to the expansion of the yeast-specific C6 transcription factor family involved in metabolic regulation.

While we have illustrated expansions in a subset of signal transduction molecules in the human genome compared with the other eukaryotic genomes, it should be noted that most of the protein domains are highly conserved. An interesting observation is that worms and humans have approximately the same number of both tyrosine kinases and serine/threonine kinases (Table 19). It is important to note, however, that these are merely counts of the catalytic domain; the proteins that contain these domains also display a wide repertoire of interaction domains with significant combinatorial diversity.

Hemostasis. Hemostasis is regulated primarily by plasma proteases of the coagulation pathway and by the interactions that occur between the vascular endothelium and platelets. Consistent with known anatomical and physiological differences between vertebrates and invertebrates, extracellular adhesion domains that constitute proteins integral to hemostasis are expanded in the human relative to the fly and worm (Tables 18 and 19). We note the evolution of domains such as FIMAC, FN1, FN2, and C1q that mediate surface interactions between hematopoeitic cells and the vascular matrix. In addition, there has been extensive recruitment of more-ancient animal-specific domains such as VWA, VWC, VWD, kringle, and FN3 into multidomain proteins that are involved in hemostatic regulation. Although we do not find a large expansion in the total number of serine proteases, this enzymatic domain has been specifically recruited into several of these multidomain proteins for proteolytic regulation in the vascular compartment. These are represented in plasma proteins that belong to the kinin and complement pathways. There is a

significant expansion in two families of matrix metalloproteases: ADAM (a disintegrin and metalloprotease) and MMPs (matrix metalloproteases) (Table 19). Proteolysis of extracellular matrix (ECM) proteins is critical for tissue development and for tissue degradation in diseases such as cancer, arthritis, Alzheimer's disease, and a variety of inflammatory conditions (*135*, *136*). ADAMs are a family of integral membrane proteins with a pivotal role in fibrinogenolysis and modulating interactions between hematopoietic components and the vascular matrix components. These proteins have been shown to cleave matrix proteins, and even signaling molecules: ADAM-17 converts tumor necrosis factor–α, and ADAM-10 has been implicated in the Notch signaling pathway (*135*). We have identified 19 members of the matrix metalloprotease family, and a total of 51 members of the ADAM and ADAM-TS families.

Apoptosis. Evolutionary conservation of some of the apoptotic pathway components across eukarya is consistent with its central role in developmental regulation and as a response to pathogens and stress signals. The signal transduction pathways involved in programmed cell death, or apoptosis, are mediated by interactions between well-characterized domains that include extracellular domains, adaptor (protein-protein interaction) domains, and those found in effector and regulatory enzymes (*137*). We enumerated the protein counts of central adaptor and effector enzyme domains that are found only in the apoptotic pathways to provide an estimate of divergence across eukarya and relative expansion in the human genome when compared with the fly and worm (Table 18). Adaptor domains found in proteins restricted only to apoptotic regulation such as the DED domains are vertebrate-specific, whereas others like BIR, CARD, and Bcl2 are represented in the fly and worm (although the number of Bcl2 family members in humans is significantly expanded). Although plants and yeast lack the caspases, caspase-like molecules, namely the para- and meta-caspases, have been reported in these organisms (*138*). Compared with other animal genomes, the human genome shows an expansion in the adaptor and effector domain–containing proteins involved in apoptosis, as well as in the proteases involved in the cascade such as the caspase and calpain families.

Expansions of other protein families. *Metabolic enzymes.* There are fewer cytochrome P450 genes in humans than in either the fly or worm. Lipoxygenases (six in humans), on the other hand, appear to be specific to the vertebrates and plants, whereas the lipoxygenase-activating proteins (four in humans) may be vertebrate-specific. Lipoxygenases are involved in arachidonic acid metabolism, and both and their activators have been implicated in diverse human pathology ranging from allergic responses to cancers. One of the most surprising human expansions, however, is in the number of glyceraldehyde-3-phosphate dehydrogenase (GAPDH) genes (46 in humans, 3 in the fly, and 4 in the worm). There is, however, evidence for many retrotransposed GAPDH pseudogenes (*139*), which may account for this apparent expansion. However, it is interesting that GAPDH, long known as a conserved enzyme involved in basic metabolism found across all phyla from bacteria to humans, has recently been shown to have other functions. It has a second cat-

Table 19. Number of proteins assigned to selected Panther families or subfamilies in *H. sapiens* (H), *D. melanogaster* (F), *C. elegans* (W), *S. cerevisiae* (Y), and *A. thaliana* (A).

Panther family/subfamily*	H	F	W	Y	A
Neural structure, function, development					
Ependymin	1	0	0	0	0
Ion channels					
Acetylcholine receptor	17	12	56	0	0
Amiloride-sensitive/degenerin	11	24	27	0	0
CNG/EAG	22	9	9	0	30
IRK	16	3	3	0	0
ITP/ryanodine	10	2	4	0	0
Neurotransmitter-gated	61	51	59	0	19
P2X purinoceptor	10	0	0	0	0
TASK	12	12	48	1	5
Transient receptor	15	3	3	1	0
Voltage-gated Ca^{2+} alpha	22	4	8	2	2
Voltage-gated Ca^{2+} alpha-2	10	3	2	0	0
Voltage-gated Ca^{2+} beta	5	2	2	0	0
Voltage-gated Ca^{2+} gamma	1	0	0	0	0
Voltage-gated K^+ alpha	33	5	11	0	0
Voltage-gated KQT	6	2	3	0	0
Voltage-gated Na^+	11	4	4	9	1
Myelin basic protein	1	0	0	0	0
Myelin PO	5	0	0	0	0
Myelin proteolipid	3	1	0	0	0
Myelin-oligodendrocyte glycoprotein	1	0	0	0	0
Neuropilin	2	0	0	0	0
Plexin	9	2	0	0	0
Semaphorin	22	6	2	0	0
Synaptotagmin	10	3	3	0	0
Immune response					
Defensin	3	0	0	0	0
Cytokine†	86	14	1	0	0
GCSF	1	0	0	0	0
GMCSF	1	0	0	0	0
Intercrine alpha	15	0	0	0	0
Intercrine beta	5	0	0	0	0
Inteferon	8	0	0	0	0
Interleukin	26	1	1	0	0
Leukemia inhibitory factor	1	0	0	0	0
MCSF	1	0	0	0	0
Peptidoglycan recognition protein	2	13	0	0	0
Pre-B cell enhancing factor	1	0	0	0	0
Small Inducible cytokine A	14	0	0	0	0
SI cytokine	2	0	0	0	0
TNF	9	0	0	0	0
Cytokine receptor†	62	1	0	0	0
Bradykinin/C-C chemokine receptor	7	0	0	0	0
FI cytokine receptor	2	0	0	0	0
Interferon receptor	3	0	0	0	0
Interleukin receptor	32	0	0	0	0
Leukocyte tyrosine kinase receptor	3	0	0	0	0
MCSF receptor	1	0	0	0	0
TNF receptor	3	0	0	0	0
Immunoglobulin receptor†	59	0	0	0	0
T-cell receptor alpha chain	16	0	0	0	0
T-cell receptor beta chain	15	0	0	0	0
T-cell receptor gamma chain	1	0	0	0	0
T-cell receptor delta chain	1	0	0	0	0
Immunoglobulin FC receptor	8	0	0	0	0
Killer cell receptor	16	0	0	0	0
Polymeric-Immunoglobulin receptor	4	0	0	0	0

alytic activity, as a uracil DNA glycosylase (*140*) and functions as a cell cycle regulator (*141*) and has even been implicated in apoptosis (*142*).

Translation. Another striking set of human expansions has occurred in certain families involved in the translational machinery. We identified 28 different ribosomal subunits that each have at least 10 copies in the genome; on average, for all ribosomal proteins there is about an 8- to 10-fold expansion in the number of genes relative to either the worm or fly. Retrotransposed pseudogenes may account for many of these expansions [see the discussion above and (*143*)]. Recent evidence suggests that a number of ribosomal proteins have secondary functions independent of their involvement in protein biosynthesis; for example, L13a and the related L7 subunits (36 copies in humans) have been shown to induce apoptosis (*144*).

There is also a four- to fivefold expansion in the elongation factor 1-alpha family (eEF1A; 56 human genes). Many of these expansions likely represent intronless paralogs that have presumably arisen from retrotransposition, and again there is evidence that many of these may be pseudogenes (*145*). However, a second form (eEF1A2) of this factor has been identied with tissue-specific expression in skeletal muscle and a complementary expression pattern to the ubiquitously expressed eEF1A (*146*).

Ribonucleoproteins. Alternative splicing results in multiple transcripts from a single gene, and can therefore generate additional diversity in an organism's protein complement. We have identified 269 genes for ribonucleoproteins. This represents over 2.5 times the number of ribonucleoprotein genes in the worm, two times that of the fly, and about the same as the 265 identified in the *Arabidopsis* genome. Whether the diversity of ribonucleoprotein genes in humans contributes to gene regulation at either the splicing or translational level is unknown.

Posttranslational modifications. In this set of processes, the most prominent expansion is the transglutaminases, calcium-dependent enzymes that catalyze the cross-linking of proteins in cellular processes such as hemostasis and apoptosis (*147*). The vitamin K–dependent gamma carboxylase gene product acts on the GLA domain (missing in the fly and worm) found in coagulation factors, osteocalcin, and matrix GLA protein (*148*). Tyrosylprotein sulfotransferases participate in the posttranslational modification of proteins involved in inflammation and hemostasis, including coagulation factors and chemokine receptors (*149*). Although there is no significant numerical increase in the counts for domains involved in nuclear protein modification, there are a number of domain arrangements in the predicted human proteins that are not found in the other currently sequenced genomes. These include the tandem association of two histone deacetylase domains in HD6 with a ubiquitin finger domain, a feature lacking in the fly genome. An additional example is the co-occurrence of important nuclear regulatory enzyme PARP (poly-ADP ribosyl transferase) domain fused to protein-interaction domains—BRCT and VWA in humans.

Table 19 (*Continued*)

Panther family/subfamily*	H	F	W	Y	A
MHC class I	22	0	0	0	0
MHC class II	20	0	0	0	0
Other Immunoglobulin†	114	0	0	0	0
Toll receptor–related	10	6	0	0	0
Developmental and homeostatic regulators					
Signaling molecules†					
Calcitonin	3	0	0	0	0
Ephrin	8	2	4	0	0
FGF	24	1	1	0	0
Glucagon	4	0	0	0	0
Glycoprotein hormone beta chain	2	0	0	0	0
Insulin	1	0	0	0	0
Insulin-like hormone	3	0	0	0	0
Nerve growth factor	3	0	0	0	0
Neuregulin/heregulin	6	0	0	0	0
neuropeptide Y	4	0	0	0	0
PDGF	1	1	0	0	0
Relaxin	3	0	0	0	0
Stannocalcin	2	0	0	0	0
Thymopoeitin	2	0	1	0	0
Thyomosin beta	4	2	0	0	0
TGF-β	29	6	4	0	0
VEGF	4	0	0	0	0
Wnt	18	6	5	0	0
Receptors†					
Ephrin receptor	12	2	1	0	0
FGF receptor	4	4	0	0	0
Frizzled receptor	12	6	5	0	0
Parathyroid hormone receptor	2	0	0	0	0
VEGF receptor	5	0	0	0	0
BDNF/NT-3 nerve growth factor receptor	4	0	0	0	0
Kinases and phosphatases					
Dual-specificity protein phosphatase	29	8	10	4	11
S/T and dual-specificity protein kinase†	395	198	315	114	1102
S/T protein phosphatase	15	19	51	13	29
Y protein kinase†	106	47	100	5	16
Y protein phosphatase	56	22	95	5	6
Signal transduction					
ARF family	55	29	27	12	45
Cyclic nucleotide phosphodiesterase	25	8	6	1	0
G protein-coupled receptors†‡	616	146	284	0	1
G-protein alpha	27	10	22	2	5
G-protein beta	5	3	2	1	1
G-protein gamma	13	2	2	0	0
Ras superfamily	141	64	62	26	86
G-protein modulators†					
ARF GTPase-activating	20	8	9	5	15
Neurofibromin	7	2	0	2	0
Ras GTPase-activating	9	3	8	1	0
Tuberin	7	3	2	0	0
Vav proto-oncogene family	35	15	13	3	0

Concluding remarks. There are several possible explanations for the differences in phenotypic complexity observed in humans when compared to the fly and worm. Some of these relate to the prominent differences in the immune system, hemostasis, neuronal, vascular, and cytoskeletal complexity. The finding that the human genome contains fewer genes than previously predicted might be compensated for by combinatorial diversity generated at the levels of protein architecture, transcriptional and translational control, posttranslational modification of proteins, or posttranscriptional regulation. Extensive domain shuffling to increase or alter combinatorial diversity can provide an exponential

increase in the ability to mediate protein-protein interactions without dramatically increasing the absolute size of the protein complement (*150*). Evolution of apparently new (from the perspective of sequence analysis) protein domains and increasing regulatory complexity by domain accretion both quantitatively and qualitatively (recruitment of novel domains with preexisting ones) are two features that we observe in humans. Perhaps the best illustration of this trend is the C2H2 zinc finger–containing transcription factors, where we see expansion in the number of domains per protein, together with vertebrate-specific domains such as KRAB and SCAN. Recent reports on the prominent use of internal ribosomal entry sites in the human genome to regulate translation of specific classes of proteins suggests that this is an area that needs further research to identify the full extent of this process in the human genome (*151*). At the posttranslational level, although we provide examples of expansions of some protein families involved in these modifications, further experimental evidence is required to evaluate whether this is correlated with increased complexity in protein processing. Posttranscriptional processing and the extent of isoform generation in the human remain to be cataloged in their entirety. Given the conserved nature of the spliceosomal machinery, further analysis will be required to dissect regulation at this level.

8 Conclusions

8.1 The whole-genome sequencing approach versus BAC by BAC

Experience in applying the whole-genome shotgun sequencing approach to a diverse group of organisms with a wide range of genome sizes and repeat content allows us to assess its strengths and weaknesses. With the success of the method for a large number of microbial genomes, *Drosophila*, and now the human, there can be no doubt concerning the utility of this method. The large number of microbial genomes that have been sequenced by this method (*15*, *80*, *152*) demonstrate that megabase-sized genomes can be sequenced efficiently without any input other that the de novo mate-paired sequences. With more complex genomes like those of *Drosophila* or human, map information, in the form of well-ordered markers, has been critical for long-range ordering of scaffolds. For joining scaffolds into chromosomes, the quality of the map (in terms of the order of the markers) is more important than the number of markers per se. Although this mapping could have been performed concurrently with sequencing, the prior existence of mapping data was beneficial. During the sequencing of the *A. thaliana* genome, sequencing of individual BAC clones permitted extension of the se-

Table 19 (*Continued*)

Panther family/subfamily*	H	F	W	Y	A
Transcription factors/chromatin organization					
C2H2 zinc finger–containing†	607	232	79	28	8
COE	7	1	1	0	0
CREB	7	1	2	0	0
ETS-related	25	8	10	0	0
Forkhead-related	34	19	15	4	0
FOS	8	2	1	0	0
Groucho	13	2	1	0	0
Histone H1	5	0	1	0	0
Histone H2A	24	1	17	3	13
Histone H2B	21	1	17	2	12
Histone H3	28	2	24	2	16
Histone H4	9	1	16	1	8
Homeotic†	168	104	74	4	78
ABD-B	5	0	0	0	0
Bithoraxoid	1	8	1	0	0
Iroquois class	7	3	1	0	0
Distal-less	5	2	1	0	0
Engrailed	2	2	1	0	0
LIM-containing	17	8	3	0	0
MEIS/KNOX class	9	4	4	2	26
NK-3/NK-2 class	9	4	5	0	0
Paired box	38	28	23	0	2
Six	5	3	4	0	0
Leucine zipper	6	0	0	0	0
Nuclear hormone receptor†	59	25	183	1	4
Pou-related	15	5	4	1	0
Runt-related	3	4	2	0	0
ECM adhesion					
Cadherin	113	17	16	0	0
Claudin	20	0	0	0	0
Complement receptor-related	22	8	6	0	0
Connexin	14	0	0	0	0
Galectin	12	5	22	0	0
Glypican	13	2	1	0	0
ICAM	6	0	0	0	0
Integrin alpha	24	7	4	0	1
Integrin beta	9	2	2	0	0
LDL receptor family	26	19	20	0	2
Proteoglycans	22	9	7	0	5
Apoptosis					
Bcl-2	12	1	0	0	0
Calpain	22	4	11	1	3
Calpain inhibitor	4	0	0	0	1
Caspase	13	7	3	0	0
Hemostasis					
ADAM/ADAMTS	51	9	12	0	0
Fibronectin	3	0	0	0	0
Globin	10	2	3	0	3
Matrix metalloprotease	19	2	7	0	3
Serum amyloid A	4	0	0	0	0
Serum amyloid P (subfamily of Pentaxin)	2	0	0	0	0
Serum paraoxonase/arylesterase	4	0	3	0	0
Serum albumin	4	0	0	0	0
Transglutaminase	10	1	0	0	0
Other enzymes					
Cytochrome p450	60	89	83	3	256
GAPDH	46	3	4	3	8
Heparan sulfotransferase	11	4	2	0	0
Splicing and translation					
EF-1alpha	56	13	10	6	13
Ribonucleoproteins†	269	135	104	60	265
Ribosomal proteins†	812	111	80	117	256

*The table lists Panther families or subfamilies relevant to the text that either (i) are not specifically represented by Pfam (Table 18) or (ii) differ in counts from the corresponding Pfam models. †This class represents a number of different families in the same Panther molecular function subcategory. ‡This count includes only rhodopsin-class, secretin-class, and metabotropic glutamate-class GPCRs.

quence well into centromeric regions and allowed high-quality resolution of complex repeat regions. Likewise, in *Drosophila*, the BAC physical map was most useful in regions near the highly repetitive centromeres and telomeres. WGA has been found to deliver excellent-quality reconstructions of the unique regions of the genome. As the genome size, and more importantly the repetitive content, increases, the WGA approach delivers less of the repetitive sequence.

The cost and overall efficiency of clone-by-clone approaches makes them difficult to justify as a stand-alone strategy for future large-scale genome-sequencing projects. Specific applications of BAC-based or other clone mapping and sequencing strategies to resolve ambiguities in sequence assembly that cannot be efficiently resolved with computational approaches alone are clearly worth exploring. Hybrid approaches to whole-genome sequencing will only work if there is sufficient coverage in both the whole-genome shotgun phase and the BAC clone sequencing phase. Our experience with human genome assembly suggests that this will require at least 3× coverage of both whole-genome and BAC shotgun sequence data.

8.2 The low gene number in humans

We have sequenced and assembled ~95% of the euchromatic sequence of *H. sapiens* and used a new automated gene prediction method to produce a preliminary catalog of the human genes. This has provided a major surprise: We have found far fewer genes (26,000 to 38,000) than the earlier molecular predictions (50,000 to over 140,000). Whatever the reasons for this current disparity, only detailed annotation, comparative genomics (particularly using the *Mus musculus* genome), and careful molecular dissection of complex phenotypes will clarify this critical issue of the basic "parts list" of our genome. Certainly, the analysis is still incomplete and considerable refinement will occur in the years to come as the precise structure of each transcription unit is evaluated. A good place to start is to determine why the gene estimates derived from EST data are so discordant with our predictions. It is likely that the following contribute to an inflated gene number derived from ESTs: the variable lengths of 3′- and 5′-untranslated leaders and trailers; the little-understood vagaries of RNA processing that often leave intronic regions in an unspliced condition; the finding that nearly 40% of human genes are alternatively spliced (*153*); and finally, the unsolved technical problems in EST library construction where contamination from heterogeneous nuclear RNA and genomic DNA are not uncommon. Of course, it is possible that there are genes that remain unpredicted owing to the absence of EST or protein data to support them, although our use of mouse genome data for predicting genes should limit this number. As was true at the beginning of genome sequencing, ultimately it will be necessary to measure mRNA in specific cell types to demonstrate the presence of a gene.

J. B. S. Haldane speculated in 1937 that a population of organisms might have to pay a price for the number of genes it can possibly carry. He theorized that when the number of genes becomes too large, each zygote carries so many new deleterious mutations that the population simply cannot maintain itself. On the basis of this premise, and on the basis of available mutation rates and x-ray–induced mutations at specific loci, Muller, in 1967 (*154*), calculated that the mammalian genome would contain a maximum of not much more than 30,000 genes (*155*). An estimate of 30,000 gene loci for humans was also arrived at by Crow and Kimura (*156*). Muller's estimate for *D. melanogaster* was 10,000 genes, compared to 13,000 derived by annotation of the fly genome (*26*, *27*). These arguments for the theoretical maximum gene number were based on simplified ideas of genetic load—that all genes have a certain low rate of mutation to a deleterious state. However, it is clear that many mouse, fly, worm, and yeast knockout mutations lead to almost no discernible phenotypic perturbations.

The modest number of human genes means that we must look elsewhere for the mechanisms that generate the complexities inherent in human development and the sophisticated signaling systems that maintain homeostasis. There are a large number of ways in which the functions of individual genes and gene products are regulated. The degree of "openness" of chromatin structure and hence transcriptional activity is regulated by protein complexes that involve histone and DNA enzymatic modifications. We enumerate many of the proteins that are likely involved in nuclear regulation in Table 19. The location, timing, and quantity of transcription are intimately linked to nuclear signal transduction events as well as by the tissue-specific expression of many of these proteins. Equally important are regulatory DNA elements that include insulators, repeats, and endogenous viruses (*157*); methylation of CpG islands in imprinting (*158*); and promoter-enhancer and intronic regions that modulate transcription. The spliceosomal machinery consists of multisubunit proteins (Table 19) as well as structural and catalytic RNA elements (*159*) that regulate transcript structure through alternative start and termination sites and splicing. Hence, there is a need to study different classes of RNA molecules (*160*) such as small nucleolar RNAs, antisense riboregulator RNA, RNA involved in X-dosage compensation, and other structural RNAs to appreciate their precise role in regulating gene expression. The phenomenon of RNA editing in which coding changes occur directly at the level of mRNA is of clinical and biological relevance (*161*). Finally, examples of translational control include internal ribosomal entry sites that are found in proteins involved in cell cycle regulation and apoptosis (*162*). At the protein level, minor alterations in the nature of protein-protein interactions, protein modifications, and localization can have dramatic effects on cellular physiology (*163*). This dynamic system therefore has many ways to modulate activity, which suggests that definition of complex systems by analysis of single genes is unlikely to be entirely successful.

In situ studies have shown that the human genome is asymmetrically populated with G+C content, CpG islands, and genes (*68*). However, the genes are not distributed quite as unequally as had been predicted (Table 9) (*69*). The most G+C-rich fraction of the genome, H3 isochores, constitute more of the genome than previously thought (about 9%), and are the most gene-dense fraction, but contain only 25% of the genes, rather than the predicted ~40%. The low G+C L isochores make up 65% of the genome, and 48% of the genes. This inhomogeneity, the net result of millions of years of mammalian gene duplication, has been described as the "desertification" of the vertebrate genome (*71*). Why are there clustered regions of high and low gene density, and are these accidents of history or driven by selection and evolution? If these deserts are dispensable, it ought to be possible to find mammalian genomes that are far smaller in size than the human genome. Indeed, many species of bats have genome sizes that are much smaller than that of humans; for example, *Miniopterus*, a species of Italian bat, has a genome size that is only 50% that of humans (*164*). Similarly, *Muntiacus*, a species of Asian barking deer, has a genome size that is ~70% that of humans.

8.3 Human DNA sequence variation and its distribution across the genome

This is the first eukaryotic genome in which a nearly uniform ascertainment of polymorphism has been completed. Although we have identified and mapped more than 3 million SNPs, this by no means implies that the task of finding and cataloging SNPs is complete. These represent only a fraction of the SNPs present in the human population as a whole. Nevertheless, this first glimpse at genome-wide variation has revealed strong inhomogeneities in the distribution of SNPs across the genome. Polymorphism in DNA carries with it a snapshot of the past operation of population genetic forces, including mutation, migration, selection, and genetic drift. The availability of a dense array of SNPs will allow questions related to each of these factors to be addressed on a genome-wide basis. SNP studies can establish the range of haplo-

types present in subjects of different ethnogeographic origins, providing insights into population history and migration patterns. Although such studies have suggested that modern human lineages derive from Africa, many important questions regarding human origins remain unanswered, and more analyses using detailed SNP maps will be needed to settle these controversies. In addition to providing evidence for population expansions, migration, and admixture, SNPs can serve as markers for the extent of evolutionary constraint acting on particular genes. The correlation between patterns of intraspecies and interspecies genetic variation may prove to be especially informative to identify sites of reduced genetic diversity that may mark loci where sequence variations are not tolerated.

The remarkable heterogeneity in SNP density implies that there are a variety of forces acting on polymorphism—sparse regions may have lower SNP density because the mutation rate is lower, because most of those regions have a lower fraction of mutations that are tolerated, or because recent strong selection in favor of a newly arisen allele "swept" the linked variation out of the population (*165*). The effect of random genetic drift also varies widely across the genome. The nonrecombining portion of the Y chromosome faces the strongest pressure from random drift because there are roughly one-quarter as many Y chromosomes in the population as there are autosomal chromosomes, and the level of polymorphism on the Y is correspondingly less. Similarly, the X chromosome has a smaller effective population size than the autosomes, and its nucleotide diversity is also reduced. But even across a single autosome, the effective population size can vary because the density of deleterious mutations may vary. Regions of high density of deleterious mutations will see a greater rate of elimination by selection, and the effective population size will be smaller (*166*). As a result, the density of even completely neutral SNPs will be lower in such regions. There is a large literature on the association between SNP density and local recombination rates in *Drosophila*, and it remains an important task to assess the strength of this association in the human genome, because of its impact on the design of local SNP densities for disease-association studies. It also remains an important task to validate SNPs on a genomic scale in order to assess the degree of heterogeneity among geographic and ethnic populations.

8.4 Genome complexity

We will soon be in a position to move away from the cataloging of individual components of the system, and beyond the simplistic notions of "this binds to that, which then docks on this, and then the complex moves there. . . ." (*167*) to the exciting area of network perturbations, nonlinear responses and thresholds, and their pivotal role in human diseases.

The enumeration of other "parts lists" reveals that in organisms with complex nervous systems, neither gene number, neuron number, nor number of cell types correlates in any meaningful manner with even simplistic measures of structural or behavioral complexity. Nor would they be expected to; this is the realm of nonlinearities and epigenesis (*168*). The 520 million neurons of the common octopus exceed the neuronal number in the brain of a mouse by an order of magnitude. It is apparent from a comparison of genomic data on the mouse and human, and from comparative mammalian neuroanatomy (*169*), that the morphological and behavioral diversity found in mammals is underpinned by a similar gene repertoire and similar neuroanatomies. For example, when one compares a pygmy marmoset (which is only 4 inches tall and weighs about 6 ounces) to a chimpanzee, the brain volume of this minute primate is found to be only about 1.5 cm^3, two orders of magnitude less than that of a chimp and three orders less than that of humans. Yet the neuroanatomies of all three brains are strikingly similar, and the behavioral characteristics of the pygmy marmoset are little different from those of chimpanzees. Between humans and chimpanzees, the gene number, gene structures and functions, chromosomal and genomic organizations, and cell types and neuroanatomies are almost indistinguishable, yet the developmental modifications that predisposed human lineages to cortical expansion and development of the larynx, giving rise to language, culminated in a massive singularity that by even the simplest of criteria made humans more complex in a behavioral sense.

Simple examination of the number of neurons, cell types, or genes or of the genome size does not alone account for the differences in complexity that we observe. Rather, it is the interactions within and among these sets that result in such great variation. In addition, it is possible that there are "special cases" of regulatory gene networks that have a disproportionate effect on the overall system. We have presented several examples of "regulatory genes" that are significantly increased in the human genome compared with the fly and worm. These include extracellular ligands and their cognate receptors (e.g., wnt, frizzled, TGF-β, ephrins, and connexins), as well as nuclear regulators (e.g., the KRAB and homeodomain transcription factor families), where a few proteins control broad developmental processes. The answers to these "complexities" perhaps lie in these expanded gene families and differences in the regulatory control of ancient genes, proteins, pathways, and cells.

8.5 Beyond single components

While few would disagree with the intuitive conclusion that Einstein's brain was more complex than that of *Drosophila*, closer comparisons such as whether the set of predicted human proteins is more complex than the protein set of *Drosophila*, and if so, to what degree, are not straightforward, since protein, protein domain, or protein-protein interaction measures do not capture context-dependent interactions that underpin the dynamics underlying phenotype.

Currently, there are more than 30 different mathematical descriptions of complexity (*170*). However, we have yet to understand the mathematical dependency relating the number of genes with organism complexity. One pragmatic approach to the analysis of biological systems, which are composed of nonidentical elements (proteins, protein complexes, interacting cell types, and interacting neuronal populations), is through graph theory (*171*). The elements of the system can be represented by the vertices of complex topographies, with the edges representing the interactions between them. Examination of large networks reveals that they can self-organize, but more important, they can be particularly robust. This robustness is not due to redundancy, but is a property of inhomogeneously wired networks. The error tolerance of such networks comes with a price; they are vulnerable to the selection or removal of a few nodes that contribute disproportionately to network stability. Gene knockouts provide an illustration. Some knockouts may have minor effects, whereas others have catastrophic effects on the system. In the case of vimentin, a supposedly critical component of the cytoplasmic intermediate filament network of mammals, the knockout of the gene in mice reveals them to be reproductively normal, with no obvious phenotypic effects (*172*), and yet the usually conspicuous vimentin network is completely absent. On the other hand, ~30% of knockouts in *Drosophila* and mice correspond to critical nodes whose reduction in gene product, or total elimination, causes the network to crash most of the time, although even in some of these cases, phenotypic normalcy ensues, given the appropriate genetic background. Thus, there are no "good" genes or "bad" genes, but only networks that exist at various levels and at different connectivities, and at different states of sensitivity to perturbation. Sophisticated mathematical analysis needs to be constantly evaluated against hard biological data sets that specifically address network dynamics. Nowhere is this more critical than in attempts to come to grips with "complexity," particularly because deconvoluting and correcting complex networks that have undergone perturbation, and have resulted in human diseases, is the greatest significant challenge now facing us.

It has been predicted for the last 15 years that complete sequencing of the human ge-

nome would open up new strategies for human biological research and would have a major impact on medicine, and through medicine and public health, on society. Effects on biomedical research are already being felt. This assembly of the human genome sequence is but a first, hesitant step on a long and exciting journey toward understanding the role of the genome in human biology. It has been possible only because of innovations in instrumentation and software that have allowed automation of almost every step of the process from DNA preparation to annotation. The next steps are clear: We must define the complexity that ensues when this relatively modest set of about 30,000 genes is expressed. The sequence provides the framework upon which all the genetics, biochemistry, physiology, and ultimately phenotype depend. It provides the boundaries for scientific inquiry. The sequence is only the first level of understanding of the genome. All genes and their control elements must be identified; their functions, in concert as well as in isolation, defined; their sequence variation worldwide described; and the relation between genome variation and specific phenotypic characteristics determined. Now we know what we have to explain.

Another paramount challenge awaits: public discussion of this information and its potential for improvement of personal health. Many diverse sources of data have shown that any two individuals are more than 99.9% identical in sequence, which means that all the glorious differences among individuals in our species that can be attributed to genes falls in a mere 0.1% of the sequence. There are two fallacies to be avoided: determinism, the idea that all characteristics of the person are "hard-wired" by the genome; and reductionism, the view that with complete knowledge of the human genome sequence, it is only a matter of time before our understanding of gene functions and interactions will provide a complete causal description of human variability. The real challenge of human biology, beyond the task of finding out how genes orchestrate the construction and maintenance of the miraculous mechanism of our bodies, will lie ahead as we seek to explain how our minds have come to organize thoughts sufficiently well to investigate our own existence.

References and Notes

1. R. L. Sinsheimer, *Genomics* 5, 954 (1989); U.S. Department of Energy, Office of Health and Environmental Research, *Sequencing the Human Genome: Summary Report of the Santa Fe Workshop*, Santa Fe, NM, 3 to 4 March 1986 (Los Alamos National Laboratory, Los Alamos, NM, 1986).
2. R. Cook-Deegan, *The Gene Wars: Science, Politics, and the Human Genome* (Norton, New York, 1996).
3. F. Sanger *et al.*, *Nature* 265, 687 (1977).
4. P. H. Seeburg *et al.*, *Trans. Assoc. Am. Physicians* 90, 109 (1977).
5. E. C. Strauss, J. A. Kobori, G. Siu, L. E. Hood, *Anal. Biochem.* 154, 353 (1986).
6. J. Gocayne *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 84, 8296 (1987).
7. A. Martin-Gallardo *et al.*, *DNA Sequence* 3, 237 (1992); W. R. McCombie *et al.*, *Nature Genet.* 1, 348 (1992); M. A. Jensen *et al.*, *DNA Sequence* 1, 233 (1991).
8. M. D. Adams *et al.*, *Science* 252, 1651 (1991).
9. M. D. Adams *et al.*, *Nature* 355, 632 (1992); M. D. Adams, A. R. Kerlavage, C. Fields, J. C. Venter, *Nature Genet.* 4, 256 (1993); M. D. Adams, M. B. Soares, A. R. Kerlavage, C. Fields, J. C. Venter, *Nature Genet.* 4, 373 (1993); M. H. Polymeropoulos *et al.*, *Nature Genet.* 4, 381 (1993); M. Marra *et al.*, *Nature Genet.* 21, 191 (1999).
10. M. D. Adams *et al.*, *Nature* 377, 3 (1995); O. White *et al.*, *Nucleic Acids Res.* 21, 3829 (1993).
11. F. Sanger, A. R. Coulson, G. F. Hong, D. F. Hill, G. B. Petersen, *J. Mol. Biol.* 162, 729 (1982).
12. B. W. J. Mahy, J. J. Esposito, J. C. Venter, *Am. Soc. Microbiol. News* 57, 577 (1991).
13. R. D. Fleischmann *et al.*, *Science* 269, 496 (1995).
14. C. M. Fraser *et al.*, *Science* 270, 397 (1995).
15. C. J. Bult *et al.*, *Science* 273, 1058 (1996); J. F. Tomb *et al.*, *Nature* 388, 539 (1997); H. P. Klenk *et al.*, *Nature* 390, 364 (1997).
16. J. C. Venter, H. O. Smith, L. Hood, *Nature* 381, 364 (1996).
17. H. Schmitt *et al.*, *Genomics* 33, 9 (1996).
18. S. Zhao *et al.*, *Genomics* 63, 321 (2000).
19. X. Lin *et al.*, *Nature* 402, 761 (1999).
20. J. L. Weber, E. W. Myers, *Genome Res.* 7, 401 (1997).
21. P. Green, *Genome Res.* 7, 410 (1997).
22. E. Pennisi, *Science* 280, 1185 (1998).
23. J. C. Venter *et al.*, *Science* 280, 1540 (1998).
24. M. D. Adams *et al.*, *Nature* 368, 474 (1994).
25. E. Marshall, E. Pennisi, *Science* 280, 994 (1998).
26. M. D. Adams *et al.*, *Science* 287, 2185 (2000).
27. G. M. Rubin *et al.*, *Science* 287, 2204 (2000).
28. E. W. Myers *et al.*, *Science* 287, 2196 (2000).
29. F. S. Collins *et al.*, *Science* 282, 682 (1998).
30. International Human Genome Sequencing Consortium (2001), *Nature* 409, 860 (2001).
31. Institutional review board: P. Calabresi (chairman), H. P. Freeman, C. McCarthy, A. L. Caplan, G. D. Rogell, J. Karp, M. K. Evans, B. Margus, C. L. Carter, R. A. Millman, S. Broder.
32. Eligibility criteria for participation in the study were as follows: prospective donors had to be 21 years of age or older, not pregnant, and capable of giving an informed consent. Donors were asked to self-define their ethnic backgrounds. Standard blood bank screens (screening for HIV, hepatitis viruses, and so forth) were performed on all samples at the clinical laboratory prior to DNA extraction in the Celera laboratory. All samples that tested positive for transmissible viruses were ineligible and were discarded. Karyotypé analysis was performed on peripheral blood lymphocytes from all samples selected for sequencing; all were normal. A two-staged consent process for prospective donors was employed. The first stage of the consent process provided information about the genome project, procedures, and risks and benefits of participating. The second stage of the consent process involved answering follow-up questions and signing consent forms, and was conducted about 48 hours after the first.
33. DNA was isolated from blood (*173*) or sperm. For sperm, a washed pellet (100 μl) was lysed in a suspension (1 ml) containing 0.1 M NaCl, 10 mM tris-Cl–20 mM EDTA (pH 8), 1% SDS, 1 mg proteinase K, and 10 mM dithiothreitol for 1 hour at 37°C. The lysate was extracted with aqueous phenol and with phenol/chloroform. The DNA was ethanol precipitated and dissolved in 1 ml TE buffer. To make genomic libraries, DNA was randomly sheared, end-polished with consecutive BAL31 nuclease and T4 DNA polymerase treatments, and size-selected by electrophoresis on 1% low-melting-point agarose. After ligation to Bst XI adapters (Invitrogen, catalog no. N408-18), DNA was purified by three rounds of gel electrophoresis to remove excess adapters, and the fragments, now with 3′-CACA overhangs, were inserted into Bst XI-linearized plasmid vector with 3′-TGTG overhangs. Libraries with three different average sizes of inserts were constructed: 2, 10, and 50 kbp. The 2-kbp fragments were cloned in a high-copy pUC18 derivative. The 10- and 50-kbp fragments were cloned in a medium-copy pBR322 derivative. The 2- and 10-kbp libraries yielded uniform-sized large colonies on plating. However, the 50-kbp libraries produced many small colonies and inserts were unstable. To remedy this, the 50-kbp libraries were digested with Bgl II, which does not cleave the vector, but generally cleaved several times within the 50-kbp insert. A 1264-bp Bam HI kanamycin resistance cassette (purified from pUCK4; Amersham Pharmacia, catalog no. 27-4958-01) was added and ligation was carried out at 37°C in the continual presence of Bgl II. As Bgl II–Bgl II ligations occurred, they were continually cleaved, whereas Bam HI–Bgl II ligations were not cleaved. A high yield of internally deleted circular library molecules was obtained in which the residual insert ends were separated by the kanamycin cassette DNA. The internally deleted libraries, when plated on agar containing ampicillin (50 μg/ml), carbenicillin (50 μg/ml), and kanamycin (15 μg/ml), produced relatively uniform large colonies. The resulting clones could be prepared for sequencing using the same procedures as clones from the 10-kbp libraries.
34. Transformed cells were plated on agar diffusion plates prepared with a fresh top layer containing no antibiotic poured on top of a previously set bottom layer containing excess antibiotic, to achieve the correct final concentration. This method of plating permitted the cells to develop antibiotic resistance before being exposed to antibiotic without the potential clone bias that can be introduced through liquid outgrowth protocols. After colonies had grown, QBot (Genetix, UK) automated colony-picking robots were used to pick colonies meeting stringent size and shape criteria and to inoculate 384-well microtiter plates containing liquid growth medium. Liquid cultures were incubated overnight, with shaking, and were scored for growth before passing to template preparation. Template DNA was extracted from liquid bacterial culture using a procedure based upon the alkaline lysis miniprep method (*173*) adapted for high throughput processing in 384-well microtiter plates. Bacterial cells were lysed; cell debris was removed by centrifugation; and plasmid DNA was recovered by isopropanol precipitation and resuspended in 10 mM tris-HCl buffer. Reagent dispensing operations were accomplished using Titertek MAP 8 liquid dispensing systems. Plate-to-plate liquid transfers were performed using Tomtec Quadra 384 Model 320 pipetting robots. All plates were tracked throughout processing by unique plate barcodes. Mated sequencing reads from opposite ends of each clone insert were obtained by preparing two 384-well cycle sequencing reaction plates from each plate of plasmid template DNA using ABI-PRISM BigDye Terminator chemistry (Applied Biosystems) and standard M13 forward and reverse primers. Sequencing reactions were prepared using the Tomtec Quadra 384-320 pipetting robot. Parent-child plate relationships and, by extension, forward-reverse sequence mate pairs were established by automated plate barcode reading by the onboard barcode reader and were recorded by direct LIMS communication. Sequencing reaction products were purified by alcohol precipitation and were dried, sealed, and stored at 4°C in the dark until needed for sequencing, at which time the reaction products were resuspended in deionized formamide and sealed immediately to prevent degradation. All sequence data were generated using a single sequencing platform, the ABI PRISM 3700 DNA Analyzer. Sample sheets were created at load time using a Java-based application that facilitates barcode scanning of the sequencing plate barcode, retrieves sample information from the central LIMS, and reserves unique trace identifiers. The application permitted a single sample sheet file in the linking directory and deleted previously created sample sheet files immediately upon scanning of a

sample plate barcode, thus enhancing sample sheet-to-plate associations.
35. F. Sanger, S. Nicklen, A. R. Coulson, *Proc. Natl. Acad. Sci. U.S.A.* 74, 5463 (1977); J. M. Prober *et al.*, *Science* 238, 336 (1987).
36. Celera's computing environment is based on Compaq Computer Corporation's Alpha system technology running the Tru64 Unix operating system. Celera uses these Alphas as Data Servers and as nodes in a Virtual Compute Farm, all of which are connected to a fully switched network operating at Fast Ethernet speed (for the VCF) and gigabit Ethernet speed (for data servers). Load balancing and scheduling software manages the submission and execution of jobs, based on central processing unit (CPU) speed, memory requirements, and priority. The Virtual Compute Farm is composed of 440 Alpha CPUs, which includes model EV6 running at a clock speed of 400 MHz and EV67 running at 667 MHz. Available memory on these systems ranges from 2 GB to 8 GB. The VCF is used to manage trace file processing, and annotation. Genome assembly was performed on a GS 160 running 16 EV67s (667 MHz) and 64 GB of memory, and 10 ES40s running 4 EV6s (500 MHz) and 32 GB of memory. A total of 100 terabytes of physical disk storage was included in a Storage Area Network that was available to systems across the environment. To ensure high availability, file and database servers were configured as 4-node Alpha TruClusters, so that services would fail over in the event of hardware or software failure. Data availability was further enhanced by using hardware- and software-based disk mirroring (RAID-0), disk striping (RAID-1), and disk striping with parity (RAID-5).
37. Trace processing generates quality values for base calls by means of Paracel's TraceTuner, trims sequence reads according to quality values, trims vector and adapter sequence from high-quality reads, and screens sequences for contaminants. Similar in design and algorithm to the phred program (*174*), TraceTuner reports quality values that reflect the log-odds score of each base being correct. Read quality was evaluated in 50-bp windows, each read being trimmed to include only those consecutive 50-bp segments with a minimum mean accuracy of 97%. End windows (both ends of the trace) of 1, 5, 10, 25, and 50 bases were trimmed to a minimum mean accuracy of 98%. Every read was further checked for vector and contaminant matches of 50 bp or more, and if found, the read was removed from consideration. Finally, any match to the 5′ vector splice junction in the initial part of a read was removed.
38. National Center for Biotechnology Information (NCBI); available at www.ncbi.nlm.nih.gov/.
39. NCBI; available at www.ncbi.nlm.nih.gov/HTGS/.
40. All bactigs over 3 kbp were examined for coverage by Celera mate pairs. An interval of a bactig was deemed an assembly error where there were no mate pairs spanning the interval and at least two reads that should have their mate on the other side of the interval but did not. In other words, there was no mate pair evidence supporting a join in the breakpoint interval and at least two mate pairs contradicting the join. By this criterion, we detected and broke apart bactigs at 13,037 locations, or equivalently, we found 2.13% of the bactigs to be misassembled.
1. We considered a BAC entry to be chimeric if, by the Lander-Waterman statistic (*175*), the odds were 0.99 or more that the assembly we produced was inconsistent with the sequence coming from a single source. By this criterion, 714 or 2.2% of BAC entries were deemed chimeric.
G. Myers, S. Selznick, Z. Zhang, W. Miller, *J. Comput. Biol.* 3, 563 (1996).
E. W. Myers, J. L. Weber, in *Computational Methods in Genome Research*, S. Suhai, Ed. (Plenum, New York, 1996), pp. 73–89.
P. Deloukas *et al.*, *Science* 282, 744 (1998).
M. A. Marra *et al.*, *Genome Res.* 7, 1072 (1997).
J. Zhang *et al.*, data not shown.
Shredded bactigs were located on long CSA scaffolds (>500 kbp) and the distribution of these fragments on the scaffolds was analyzed. If the spread of these fragments was greater than four times the reported BAC length, the BAC was considered to be chimeric. In addition, if >20% of bactigs of a given BAC were found on a different scaffolds that were not adjacent in map position, then the BAC was also considered as chimeric. The total chimeric BACs divided by the number of BACs used for CSA gave the minimal estimate of chimerism rate.
48. M. Hattori *et al.*, *Nature* 405, 311 (2000).
49. I. Dunham *et al.*, *Nature* 402, 489 (1999).
50. A. B. Carvalho, B. P. Lazzaro, A. G. Clark, *Proc. Natl. Acad. Sci. U.S.A.* 97, 13239 (2000).
51. The International RH Mapping Consortium, available at www.ncbi.nlm.nih.gov/genemap99/.
52. See http://ftp.genome.washington.edu/RM/RepeatMasker.html.
53. G. D. Schuler, *Trends Biotechnol.* 16, 456 (1998).
54. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* 215, 403 (1990).
55a. M. Olivier *et al.*, *Science* 291, 1298 (2001).
55b. See http://genome.ucsc.edu/.
56. N. Chaudhari, W. E. Hahn, *Science* 220, 924 (1983); R. J. Milner, J. G. Sutcliffe, *Nucleic Acids Res.* 11, 5497 (1983).
57. D. Dickson, *Nature* 401, 311 (1999).
58. B. Ewing, P. Green, *Nature Genet.* 25, 232 (2000).
59. H. Roest Crollius *et al.*, *Nature Genet.* 25, 235 (2000).
60. M. Yandell, in preparation.
61. K. D. Pruitt, K. S. Katz, H. Sicotte, D. R. Maglott, *Trends Genet.* 16, 44 (2000).
62. Scaffolds containing greater than 10 kbp of sequence were analyzed for features of biological importance through a series of computational steps, and the results were stored in a relational database. For scaffolds greater than one megabase, the sequence was cut into single megabase pieces before computational analysis. All sequence was masked for complex repeats using Repeatmasker (*52*) before gene finding or homology-based analysis. The computational pipeline required ~7 hours of CPU time per megabase, including repeat masking, or a total compute time of about 20,000 CPU hours. Protein searches were performed against the nonredundant protein database available at the NCBI. Nucleotide searches were performed against human, mouse, and rat Celera Gene Indices (assemblies of cDNA and EST sequences), mouse genomic DNA reads generated at Celera (3×), the Ensembl gene database available at the European Bioinformatics Institute (EBI), human and rodent (mouse and rat) EST data sets parsed from the dbEST database (NCBI), and a curated subset of the RefSeq experimental mRNA database (NCBI). Initial searches were performed on repeat-masked sequence with BLAST 2.0 (*54*) optimized for the Compaq Alpha computeserver and an effective database size of 3×10^9 for BLASTN searches and 1×10^9 for BLASTX searches. Additional processing of each query-subject pair was performed to improve the alignments. All protein BLAST results having an expectation score of $<1 \times 10^{-4}$, human nucleotide BLAST results having an expectation score of $<1 \times 10^{-8}$ with >94% identity, and rodent nucleotide BLAST results having an expectation score of $<1 \times 10^{-8}$ with >80% identity were then examined on the basis of their high-scoring pair (HSP) coordinates on the scaffold to remove redundant hits, retaining hits that supported possible alternative splicing. For BLASTX searches, analysis was performed separately for selected model organisms (yeast, mouse, human, *C. elegans*, and *D. melanogaster*) so as not to exclude HSPs from these organisms that support the same gene structure. Sequences producing BLAST hits judged to be informative, nonredundant, and sufficiently similar to the scaffold sequence were then realigned to the genomic sequence with Sim4 for ESTs, and with Lap for proteins. Because both of these algorithms take splicing into account, the resulting alignments usually give a better representation of intron-exon boundaries than standard BLAST analyses and thus facilitate further annotation (both machine and human). In addition to the homology-based analysis described above, three ab initio gene prediction programs were used (*63*).
63. E. C. Uberbacher, Y. Xu, R. J. Mural, *Methods Enzymol.* 266, 259 (1996); C. Burge, S. Karlin, *J. Mol. Biol.* 268, 78 (1997); R. J. Mural, *Methods Enzymol.* 303, 77 (1999); A. A. Salamov, V. V. Solovyev, *Genome Res.* 10, 516 (2000); Floreal *et al.*, *Genome Res.* 8, 967 (1998).
64. G. L. Miklos, B. John, *Am. J. Hum. Genet.* 31, 264 (1979); U. Francke, *Cytogenet. Cell Genet.* 65, 206 (1994).
65. P. E. Warburton, H. F. Willard, in *Human Genome Evolution*, M. S. Jackson, T. Strachan, G. Dover, Eds. (BIOS Scientific, Oxford, 1996), pp. 121–145.
66. J. E. Horvath, S. Schwartz, E. E. Eichler, *Genome Res.* 10, 839 (2000).
67. W. A. Bickmore, A. T. Sumner, *Trends Genet.* 5, 144 (1989).
68. G. P. Holmquist, *Am. J. Hum. Genet.* 51, 17 (1992).
69. G. Bernardi, *Gene* 241, 3 (2000).
70. S. Zoubak, O. Clay, G. Bernardi, *Gene* 174, 95 (1996).
71. S. Ohno, *Trends Genet.* 1, 160 (1985).
72. K. W. Broman, J. C. Murray, V. C. Sheffield, R. L. White, J. L. Weber, *Am. J. Hum. Genet.* 63, 861 (1998).
73. M. J. McEachern, A. Krauskopf, E. H. Blackburn, *Annu. Rev. Genet.* 34, 331 (2000).
74. A. Bird, *Trends Genet.* 3, 342 (1987).
75. M. Gardiner-Garden, M. Frommer, *J. Mol. Biol.* 196, 261 (1987).
76. F. Larsen, G. Gundersen, R. Lopez, H. Prydz, *Genomics* 13, 1095 (1992).
77. S. H. Cross, A. Bird, *Curr. Opin. Genet. Dev.* 5, 309 (1995).
78. J. Peters, *Genome Biol.* 1, reviews1028.1 (2000) (http://genomebiology.com/2000/1/5/reviews/1028).
79. C. Grunau, W. Hindermann, A. Rosenthal, *Hum. Mol. Genet.* 9, 2651 (2000).
80. F. Antequera, A. Bird, *Proc. Natl. Acad. Sci. U.S.A.* 90, 11995 (1993).
81. S. H. Cross *et al.*, *Mamm. Genome* 11, 373 (2000).
82. D. Slavov *et al.*, *Gene* 247, 215 (2000).
83. A. F. Smit, A. D. Riggs, *Nucleic Acids Res.* 23, 98 (1995).
84. D. J. Elliott *et al.*, *Hum. Mol. Genet.* 9, 2117 (2000).
85. A. V. Makeyev, A. N. Chkheidze, S. A. Lievhaber, *J. Biol. Chem.* 274, 24849 (1999).
86. Y. Pan, W. K. Decker, A. H. H. M. Huq, W. J. Craigen, *Genomics* 59, 282 (1999).
87. P. Nouvel, *Genetica* 93, 191 (1994).
88. I. Goncalves, L. Duret, D. Mouchiroud, *Genome Res.* 10, 672 (2000).
89. Lek first compares all proteins in the proteome to one another. Next, the resulting BLAST reports are parsed, and a graph is created wherein each protein constitutes a node; any hit between two proteins with an expectation beneath a user-specified threshold constitutes an edge. Lek then uses this graph to compute a similarity between each protein pair *ij* in the context of the graph as a whole by simply dividing the number of BLAST hits shared in common between the two proteins by the total number of proteins hit by *i* and *j*. This simple metric has several interesting properties. First, because the similarity metric takes into account both the similarity and the differences between the two sequences at the level of BLAST hits, the metric respects the multidomain nature of protein space. Two multidomain proteins, for instance, each containing domains A and B, will have a greater pairwise similarity to each other than either one will have to a protein containing only A or B domains, so long as A-B–containing multidomain proteins are less frequent in the proteome than are single-domain proteins containing A or B domains. A second interesting property of this similarity metric is that it can be used to produce a similarity matrix for the proteome as a whole without having to first produce a multiple alignment for each protein family, an error-prone and very time-consuming process. Finally, the metric does not require that either sequence have significant homology to the other in order to have a defined similarity to each other, only that they

share at least one significant BLAST hit in common. This is an especially interesting property of the metric, because it allows the rapid recovery of protein families from the proteome for which no multiple alignment is possible, thus providing a computational basis for the extension of protein homology searches beyond those of current HMM- and profile-based search methods. Once the whole-proteome similarity matrix has been calculated, Lek first partitions the proteome into single-linkage clusters (*27*) on the basis of one or more shared BLAST hits between two sequences. Next, these single-linkage clusters are further partitioned into subclusters, each member of which shares a user-specified pairwise similarity with the other members of the cluster, as described above. For the purposes of this publication, we have focused on the analysis of single-linkage clusters and what we have termed "complete clusters," e.g., those subclusters for which every member has a similarity metric of 1 to every other member of the subcluster. We believe that the single-linkage and complete clusters are of special interest, in part, because they allow us to estimate and to compare sizes of core protein sets in a rigorous manner. The rationale for this is as follows: If one imagines for a moment a perfect clustering algorithm capable of perfectly partitioning one or more perfectly annotated protein sets into protein families, it is reasonable to assume that the number of clusters will always be greater than, or equal to, the number of single-linkage clusters, because single-linkage clustering is a maximally agglomerative clustering method. Thus, if there exists a single protein in the predicted protein set containing domains A and B, then it will be clustered by single linkage together with all single-domain proteins containing domains A or B. Likewise, for a predicted protein set containing a single multidomain protein, the number of real clusters must always be less than or equal to the number of complete clusters, because it is impossible to place a unique multidomain protein into a complete cluster. Thus, the single-linkage and complete clusters plus singletons should comprise a lower and upper bound of sizes of core protein sets, respectively, allowing us to compare the relative size and complexity of different organisms' predicted protein set.

90. T. F. Smith, M. S. Waterman, *J. Mol. Biol.* 147, 195 (1981).
91. A. L. Delcher *et al.*, *Nucleic Acids Res.* 27, 2369 (1999).
92. *Arabidopsis* Genome Initiative, *Nature* 408, 796 (2000).

93. The probability that a contiguous set of proteins is the result of a segmental duplication can be estimated approximately as follows. Given that protein A and B occur on one chromosome, and that A′ and B′ (paralogs of A and B) also exist in the genome, the probability that B′ occurs immediately after A′ is $1/N$, where N is the number of proteins in the set (for this analysis, $N = 26{,}588$). Allowing for B′ to occur as any of the next J-1 proteins [leaving a gap between A′ and B′ increases the probability to $(J - 1)/N$; allowing B′A′ or A′B′ gives a probability of $2(J - 1)/N$]. Considering three genes ABC, the probability of observing A′B′C′ elsewhere in the genome, given that the paralogs exist, is $1/N^2$. Three proteins can occur across a spread of five positions in six ways; more generally, we compute the number of ways that K proteins can be spread across J positions by counting all possible arrangements of $K - 2$ proteins in the $J - 2$ positions between the first and last protein. Allowing for a spread to vary from K positions (no gaps) to J gives

$$L = \sum_{x=K-2}^{J-2} \binom{x}{K-2}$$

arrangements. Thus, the probability of chance occurrence is L/N^{K-1}. Allowing for both sets of genes (e.g., ABC and A′B′C′) to be spread across J positions increases this to L^2/N^{K-1}. The duplicated segment might be rearranged by the operations of reversal or translocation; allowing for M such rearrangements gives us a probability $P = L^2M/N^{K-1}$. For example, the probability of observing a duplicated set of three genes in two different locations, where the three genes occur across a spread of five positions in both locations, is $36/N^2$; the expected number of such matched sets in the predicted protein set is approximately $(N)36/N^2 = 36/N$, a value $\ll 1$. Therefore, any such duplications of three genes are unlikely to result from random rearrangements of the genome. If any of the genes occur in more than two copies, the probability that the apparent duplication has occurred by chance increases. The algorithm for selecting candidate duplications only generates matched protein sets with $P \ll 1$.

94. B. J. Trask *et al.*, *Hum. Mol. Genet.* 7, 13 (1998); D. Sharon *et al.*, *Genomics* 61, 24 (1999).
95. W. B. Barbazuk *et al.*, *Genome Res.* 10, 1351 (2000); A. McLysaght, A. J. Enright, L. Skrabanek, K. H. Wolfe, *Yeast* 17, 22 (2000); D. W. Burt *et al.*, *Nature* 402, 411 (1999).
96. Reviewed in L. Skrabanek, K. H. Wolfe, *Curr. Opin. Genet. Dev.* 8, 694 (1998).
97. P. Taillon-Miller, Z. Gu, Q. Li, L. Hillier, P. Y. Kwok, *Genome Res.* 8, 748 (1998); P. Taillon-Miller, E. E. Piernot, P. Y. Kwok, *Genome Res.* 9, 499 (1999).
98. D. Altshuler *et al.*, *Nature* 407, 513 (2000).
99. G. T. Marth *et al.*, *Nature Genet.* 23, 452 (1999).
100. W.-H. Li, *Molecular Evolution* (Sinauer, Sunderland, MA, 1997).
101. M. Cargill *et al.*, *Nature Genet.* 22, 231 (1999).
102. M. K. Halushka *et al.*, *Nature Genet.* 22, 239 (1999).
103. J. Zhang, T. L. Madden, *Genome Res.* 7, 649 (1997).
104. M. Nei, *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York, 1987).

105. From the observed coverage of the sequences at each site for each individual, we calculated the probability that a SNP would be detected at the site if it were present. For each level of coverage, there is a binomial sampling of the two homologs for each individual, and a heterozygous site could only be ascertained if both homologs are present, or if two alleles from different individuals are present. With coverage x from a given individual, both homologs are present in the assembly with probability $1 - (1/2)x^{-1}$. Even if both homologs are present, the probability that a SNP is detected is <1 because a fraction of sites failed the quality criteria. Integrating over coverage levels, the binomial sampling, and the quality distribution, we derived an expected number of sites in the genome that were ascertained for polymorphism for each individual. The nucleotide diversity was then the observed number of variable sites divided by the expected number of sites ascertained.

106. M. W. Nachman, V. L. Bauer, S. L. Crowell, C. F. Aquadro, *Genetics* 150, 1133 (1998).
107. D. A. Nickerson *et al.*, *Nature Genet.* 19, 233 (1998); D. A. Nickerson *et al.*, *Genomic Res.* 10, 1532 (2000); L. Jorde *et al.*, *Am. J. Hum. Genet.* 66, 979 (2000); D. G. Wang *et al.*, *Science* 280, 1077 (1998).
108. M. Przeworski, R. R. Hudson, A. Di Rienzo, *Trends Genet.* 16, 296 (2000).
109. S. Tavare, *Theor. Popul. Biol.* 26, 119 (1984).
110. R. R. Hudson, in *Oxford Surveys in Evolutionary Biology*, D. J. Futuyma, J. D. Antonovics, Eds. (Oxford Univ. Press, Oxford, 1990), vol. 7, pp. 1–44.
111. A. G. Clark *et al.*, *Am. J. Hum. Genet.* 63, 595 (1998).
112. M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, 1983).
113. H. Kaessmann, F. Heissig, A. von Haeseler, S. Paabo, *Nature Genet.* 22, 78 (1999).
114. E. L. Sonnhammer, S. R. Eddy, R. Durbin, *Proteins* 28, 405 (1997).
115. A. Bateman *et al.*, *Nucleic Acids Res.* 28, 263 (2000).

116. Brief description of the methods used to build the Panther classification. First, the June 2000 release of the GenBank NR protein database (excluding sequences annotated as fragments or mutants) was partitioned into clusters using BLASTP. For the clustering, a seed sequence was randomly chosen, and the cluster was defined as all sequences matching the seed to statistical significance (E-value $< 10^{-5}$) and "globally" alignable (the length of the match region must be >70% and <130% of the length of the seed). If the cluster had more than five members, and at least one from a multicellular eukaryote, the cluster was extended. For the extension step, a hidden Markov Model (HMM) was trained for the cluster, using the SAM software package, version 2. The HMM was then scored against GenBank NR (excluding mutants but including fragments for this step), and all sequences scoring better than a specific (NLL-NULL) score were added to the cluster. The HMM was then retrained (with fixed model length) and all sequences in the cluster were aligned to the HMM to produce a multiple sequence alignment. This alignment was assessed by a number of quality measures. If the alignment failed the quality check, the initial cluster was rebuilt around the seed using a more restrictive E-value, followed by extension, alignment, and reassessment. This process was repeated until the alignment quality was good. The multiple alignment and "general" (i.e., describing the entire cluster, or "family") HMM (*176*) were then used as input into the BETE program (*177*). BETE calculates a phylogenetic tree for the sequences in the alignment. Functional information about the sequences in each cluster were parsed from SwissProt (*178*) and GenBank records. "Tree-attribute viewer" software was used by biologist curators to correlate the phylogenetic tree with protein function. Subfamilies were manually defined on the basis of shared function across subtrees, and were named accordingly. HMMs were then built for each subfamily, using information from both the subfamily and family (K. Sjölander, in preparation). Families were also manually named according to the functions contained within them. Finally, all of the families and subfamilies were classified into categories and subcategories based on their molecular functions. The categorization was done by manual review of the family and subfamily names, by examining SwissProt and GenBank records, and by review of the literature as well as resources on the World Wide Web. The current version (2.0) of the Panther molecular function schema has four levels: category, subcategory, family, and subfamily. Protein sequences for whole eukaryotic genomes (for the predicted human proteins and annotated proteins for fly, worm, yeast, and *Arabidopsis*) were scored against the Panther library of family and subfamily HMMs. If the score was significant (the NLL-NULL score cutoff depends on the protein family), the protein was assigned to the family or subfamily function with the most significant score.

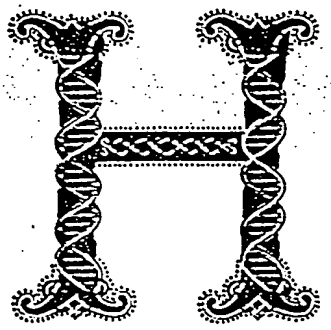
117. C. P. Ponting, J. Schultz, F. Milpetz, P. Bork, *Nucleic Acids Res.* 27, 229 (1999).
118. A. Goffeau *et al.*, *Science* 274, 546, 563 (1996).
119. *C. elegans* Sequencing Consortium, *Science* 282, 2012 (1998).
120. S. A. Chervitz *et al.*, *Science* 282, 2022 (1998).
121. E. R. Kandel, J. H. Schwartz, T. Jessell, *Principles of Neural Science* (McGraw-Hill, New York, ed. 4, 2000).
122. D. A. Goodenough, J. A. Goliger, D. L. Paul, *Annu. Rev. Biochem.* 65, 475 (1996).
123. D. G. Wilkinson, *Int. Rev. Cytol.* 196, 177 (2000).
124. F. Nakamura, R. G. Kalb, S. M. Strittmatter, *J. Neurobiol.* 44, 219 (2000).
125. P. J. Horner, F. H. Gage, *Nature* 407, 963 (2000); P. Casaccia-Bonnefil, C. Gu, M. V. Chao, *Adv. Exp. Med. Biol.* 468, 275 (1999).
126. S. Wang, B. A. Barres, *Neuron* 27, 197 (2000).
127. M. Geppert, T. C. Sudhof, *Annu. Rev. Neurosci.* 21, 75 (1998); J. T. Littleton, H. J. Bellen, *Trends Neurosci.* 18, 177 (1995).
128. A. Maximov, T. C. Sudhof, I. Bezprozvanny, *J. Biol. Chem.* 274, 24453 (1999).
129. B. Sampo *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 97, 3666 (2000).
130. G. Lemke, *Glia* 7, 263 (1993).
131. M. Bernfield *et al.*, *Annu. Rev. Biochem.* 68, 729 (1999).
132. N. Perrimon, M. Bernfield, *Nature* 404, 725 (2000).
133. U. Lindahl, M. Kusche-Gullberg, L. Kjellen, *J. Biol. Chem.* 273, 24979 (1998).
134. J. L. Riechmann *et al.*, *Science* 290, 2105 (2000).
135. T. L. Hurskainen, S. Hirohata, M. F. Seldin, S. S. Apte, *J. Biol. Chem.* 274, 25555 (1999).

136. R. A. Black, J. M. White, *Curr. Opin. Cell Biol.* 10, 654 (1998).
137. L. Aravind, V. M. Dixit, E. V. Koonin, *Trends Biochem. Sci.* 24, 47 (1999).
138. A. G. Uren *et al.*, *Mol. Cell* 6, 961 (2000).
139. P. Garcia-Meunier, M. Etienne-Julan, P. Fort, M. Piechaczyk, F. Bonhomme, *Mamm. Genome* 4, 695 (1993).
140. K. Meyer-Siegler *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 88, 8460 (1991).
141. N. R. Mansur, K. Meyer-Siegler, J. C. Wurzer, M. A. Sirover, *Nucleic Acids Res.* 21, 993 (1993).
142. N. A. Tatton, *Exp. Neurol.* 166, 29 (2000).
143. N. Kenmochi *et al.*, *Genome Res.* 8, 509 (1998).
144. F. W. Chen, Y. A. Ioannou, *Int. Rev. Immunol.* 18, 429 (1999).
145. H. O. Madsen, K. Poulsen, O. Dahl, B. F. Clark, J. P. Hjorth, *Nucleic Acids Res.* 18, 1513 (1990).
146. D. M. Chambers, J. Peters, C. M. Abbott, *Proc. Natl. Acad. Sci. U.S.A.* 95, 4463 (1998); A. Khalyfa, B. M. Carlson, J. A. Carlson, E. Wang, *Dev. Dyn.* 216, 267 (1999).
147. D. Aeschlimann, V. Thomazy, *Connect. Tissue Res.* 41, 1 (2000).
148. P. Munroe *et al.*, *Nature Genet.* 21, 142 (1999); S. M. Wu, W. F. Cheung, D. Frazier, D. W. Stafford, *Science* 254, 1634 (1991); B. Furie *et al.*, *Blood* 93, 1798 (1999).
149. J. W. Kehoe, C. R. Bertozzi, *Chem. Biol.* 7, R57 (2000).
150. T. Pawson, P. Nash, *Genes Dev.* 14, 1027 (2000).
151. A. W. van der Velden, A. A. Thomas, *Int. J. Biochem. Cell Biol.* 31, 87 (1999).
152. C. M. Fraser *et al.*, *Science* 281, 375 (1998); H. Tettelin *et al.*, *Science* 287, 1809 (2000).
153. D. Brett *et al.*, *FEBS Lett.* 474, 83 (2000).
154. H. J. Muller, H. Kern, *Z. Naturforsch. B* 22, 1330 (1967).
155. H. J. Muller, in *Heritage from Mendel*, R. A. Brink, Ed. (Univ. of Wisconsin Press, Madison, WI, 1967), p. 419.
156. J. F. Crow, M. Kimura, *Introduction to Population Genetics Theory* (Harper & Row, New York, 1970).
157. K. Kobayashi *et al.*, *Nature* 394, 388 (1998).
158. A. P. Feinberg, *Curr. Top. Microbiol. Immunol.* 249, 87 (2000).
159. C. A. Collins, C. Guthrie, *Nature Struct. Biol.* 7, 850 (2000).
160. S. R. Eddy, *Curr. Opin. Genet. Dev.* 9, 695 (1999).
161. Q. Wang, J. Khillan, P. Gadue, K. Nishikura, *Science* 290, 1765 (2000).
162. M. Holcik, N. Sonenberg, R. G. Korneluk, *Trends Genet.* 16, 469 (2000).
163. T. A. McKinsey, C. L. Zhang, J. Lu, E. N. Olson, *Nature* 408, 106 (2000).
164. E. Capanna, M. G. M. Romanini, *Caryologia* 24, 471 (1971).
165. J. Maynard Smith, *J. Theor. Biol.* 128, 247 (1987).
166. D. Charlesworth, B. Charlesworth, M. T. Morgan, *Genetics* 141, 1619 (1995).
167. J. E. Bailey, *Nature Biotechnol.* 17, 616 (1999).
168. R. Maleszka, H. G. de Couet, G. L. Miklos, *Proc. Natl. Acad. Sci. U.S.A.* 95, 3731 (1998).
169. G. L. Miklos, *J. Neurobiol.* 24, 842 (1993).
170. J. P. Crutchfield, K. Young, *Phys. Rev. Lett.* 63, 105 (1989); M. Gell-Mann, S. Lloyd, *Complexity* 2, 44 (1996).
171. A. L. Barabasi, R. Albert, *Science* 286, 509 (1999).
172. E. Colucci-Guyon *et al.*, *Cell* 79, 679 (1994).
173. J. Sambrook, E. F. Fritch, T. Maniatis, *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, ed. 2, 1989).
174. B. Ewing, P. Green, *Genome Res.* 8, 186 (1998); B. Ewing, L. Hillier, M. C. Wendl, P. Green, *Genome Res.* 8, 175 (1998).
175. E. S. Lander, M. S. Waterman, *Genomics* 2, 231 (1988).
176. A. Krogh, K. Sjölander, *J. Mol. Biol.* 235, 1501 (1994).
177. K. Sjölander, *Proc. Int. Soc. Mol. Biol.* 6, 165 (1998).
178. A. Bairoch, R. Apweiler, *Nucleic Acids Res.* 28, 45 (2000).
179. GO, available at www.geneontology.org/.
180. R. L. Tatusov, M. Y. Galperin, D. A. Natale, E. V. Koonin, *Nucleic Acids Res.* 28, 33 (2000).
181. We thank E. Eichler and J. L. Goldstein for many helpful discussions and critical reading of the manuscript, and A. Caplan for advice and encouragement. We also thank T. Hein, D. Lucas, G. Edwards, and the Celera IT staff for outstanding computational support. The cost of this project was underwritten by the Celera Genomics Group of the Applera Corporation. We thank the Board of Directors of Applera Corporation: J. F. Abely Jr. (retired), R. H. Ayers, J.-L. Bélingard, R. H. Hayes, A. J. Levine, T. E. Martin, C. W. Slayman, O. R. Smith, G. C. St. Laurent Jr., and J. R. Tobin for their vision, enthusiasm, and unwavering support and T. L. White for leadership and advice. Data availability: The genome sequence and additional supporting information are available to academic scientists at the Web site (www.celera.com). Instructions for obtaining a DVD of the genome sequence can be obtained through the Web site. For commercial scientists wishing to verify the results presented here, the genome data are available upon signing a Material Transfer Agreement, which can also be found on the Web site.

5 December 2000; accepted 19 January 2001

EXHIBIT T

THE HUMAN GENOME



Humanity has been given a great gift. With the completion of the human genome sequence, we have received a powerful tool for unlocking the secrets of our genetic heritage and for finding our place among the other participants in the adventure of life.

This week's issue of *Science* contains the report of the sequencing of the human genome from a group of authors led by Craig Venter of Celera Genomics. The report of the sequencing of the human genome from the publicly funded consortium of laboratories led by Francis Collins appears in this week's *Nature*. This stunning achievement has been portrayed—often unfairly—as a competition between two ventures, one public and one private. That characterization detracts from the awesome accomplishment jointly unveiled this week. In truth, each project contributed to the other. The inspired vision that launched the publicly funded project roughly 10 years ago reflected, and now rewards, the confidence of those who believe that the pursuit of large-scale fundamental problems in the life sciences is in the national interest. The technical innovation and drive of Craig Venter and his colleagues made it possible to celebrate this accomplishment far sooner than was believed possible. Thus, we can salute what has become, in the end, not a contest but a marriage (perhaps encouraged by shotgun) between public funding and private entrepreneurship.

A historic moment for the scientific endeavor.

There are excellent scientific reasons for applauding an outcome that has given us two winners. Two sequences are better than one; the opportunity for comparison and convergence is invaluable. Indeed, a real-world proof of the importance of access to both sets of data can be found in the pages of this issue of *Science*, in the comparative analysis by Olivier *et al.* (p. 1298).

Although we have made the point before, it is worth repeating that the sequencing of the human genome represents, not an ending, but the beginning of a new approach to biology. As Galas says in his Viewpoint (p. 1257), the knowledge that all of the genetic components of any process can be identified will give extraordinary new power to scientists. Because of this breakthrough, research can evolve from analyzing the effects of individual genes to a more integrated view that examines whole ensembles of genes as they interact to form a living human being. Several articles in this issue highlight how this approach is already beginning to revolutionize the way we look at human disease.

This has been a massive project, on a scale unparalleled in the history of biology, but of course it has built on the scientific insights of centuries of investigators. By coincidence, this landmark announcement falls during the week of the anniversary of the birth of Charles Darwin. Darwin's message that the survival of a species can depend on its ability to evolve in the face of change is peculiarly pertinent to discussions that have gone on in the past year over access to the Celera data. (Full information regarding the agreements that were reached to make the data available can be found at www.sciencemag.org/feature/data/announcement/gsp.shl.) We are willing to be flexible in allowing data repositories other than the traditional GenBank, while insisting on access to all the data needed to verify conclusions. In this domain, change is everywhere: Commercial researchers are producing more and more potentially valuable sequences, yet (at least in the United States) laws governing databases provide scant protection against piracy. Had the Celera data been kept secret, it would have been a serious loss to the scientific community. We hope that our adaptability in the face of change will enable other proprietary data to be published after peer review, in a way that satisfies our continuing commitment to full access.

It should be no surprise that an achievement so stunning, and so carefully watched, has created new challenges for the scientific venture. *Science* is proud to have played a role in bringing this discovery onto the public stage. It is literally true that this is a historic moment for the scientific endeavor. The human genome has been called the Book of Life. Rather, it is a library, in which, with rules that encourage exploration and reward creativity, we can find many of the books that will help define us and our place in the great tapestry of life.

Barbara R. Jasny and Donald Kennedy

EXHIBIT U

Query= SEQ ID NO:1
(1278 letters)

Sequences producing significant alignments:	Score (bits)	E Value
AC025750.10.1.151367	1287	0.0

>AC025750.10.1.151367
Length = 151367

Score = 1287 bits (649), Expect = 0.0
Identities = 649/649 (100%)
Strand = Plus / Minus

```
Query: 630  caggataattgaagctatctgcataggttggttcactgccgagtgcatcgtgaggttcat 689
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 8482 caggataattgaagctatctgcataggttggttcactgccgagtgcatcgtgaggttcat 8423


Query: 690  tgtctccaaaaacaagtgtgagtttgtcaagagacccctgaacatcattgatttactggc 749
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 8422 tgtctccaaaaacaagtgtgagtttgtcaagagacccctgaacatcattgatttactggc 8363


Query: 750  aatcacgccgtattacatctctgtgttgatgacagtgtttacaggcgagaactctcaact 809
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 8362 aatcacgccgtattacatctctgtgttgatgacagtgtttacaggcgagaactctcaact 8303


Query: 810  ccagagggctggagtcaccttgagggtacttagaatgatgaggatttttgggtgattaa 869
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 8302 ccagagggctggagtcaccttgagggtacttagaatgatgaggatttttgggtgattaa 8243


Query: 870  gcttgcccgtcacttcattggtcttcagacactcggtttgactctcaaacgttgctaccg 929
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 8242 gcttgcccgtcacttcattggtcttcagacactcggtttgactctcaaacgttgctaccg 8183


Query: 930  agagatggttatgttacttgtcttcatttgtgttgccatggcaatctttagtgcactttc 989
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 8182 agagatggttatgttacttgtcttcatttgtgttgccatggcaatctttagtgcactttc 8123


Query: 990  tcagcttcttgaacatgggctggacctggaaacatccaacaaggactttaccagcattcc 1049
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 8122 tcagcttcttgaacatgggctggacctggaaacatccaacaaggactttaccagcattcc 8063


Query: 1050 tgctgcctgctggtgggtgattatctctatgactacagttggctatggagatatgtatcc 1109
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 8062 tgctgcctgctggtgggtgattatctctatgactacagttggctatggagatatgtatcc 8003
```

```
Query: 1110 tatcacagtgcctggaagaattcttggaggagtttgtgttgtcagtggaattgttctatt 1169
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 8002 tatcacagtgcctggaagaattcttggaggagtttgtgttgtcagtggaattgttctatt 7943


Query: 1170 ggcattacctatcacttttatctaccatagctttgtgcagtgttatcatgagctcaagtt 1229
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 7942 ggcattacctatcacttttatctaccatagctttgtgcagtgttatcatgagctcaagtt 7883


Query: 1230 tagatctgctaggtatagtaggagcctctccactgaattcctgaattaa 1278
           ||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 7882 tagatctgctaggtatagtaggagcctctccactgaattcctgaattaa 7834


 Score = 1255 bits (633), Expect = 0.0
 Identities = 633/633 (100%)
 Strand = Plus / Minus


Query: 1     atgaccttcgggcgcagcggggcggcctcggtggtgctgaacgtgggcggcgcccggtat 60
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 57401 atgaccttcgggcgcagcggggcggcctcggtggtgctgaacgtgggcggcgcccggtat 57342


Query: 61    tcgctgtcccgggagctgctgaaggacttcccgctgcgccgcgtgagccggctgcacggc 120
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 57341 tcgctgtcccgggagctgctgaaggacttcccgctgcgccgcgtgagccggctgcacggc 57282


Query: 121   tgccgctccgagcgcgacgtgctcgaggtgtgcgacgactacgaccgcgagcgcaacgag 180
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 57281 tgccgctccgagcgcgacgtgctcgaggtgtgcgacgactacgaccgcgagcgcaacgag 57222


Query: 181   tacttcttcgaccggcactcggaggccttcggcttcatcctgctctacgtgcgcggccac 240
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 57221 tacttcttcgaccggcactcggaggccttcggcttcatcctgctctacgtgcgcggccac 57162


Query: 241   ggcaagctgcgcttcgcgccgcggatgtgcgagctctccttctacaacgagatgatctac 300
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 57161 ggcaagctgcgcttcgcgccgcggatgtgcgagctctccttctacaacgagatgatctac 57102


Query: 301   tggggcctggagggcgcgcacctcgagtactgctgccagcgccgcctcgacgaccgcatg 360
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 57101 tggggcctggagggcgcgcacctcgagtactgctgccagcgccgcctcgacgaccgcatg 57042


Query: 361   tccgacacctacaccttctactcggccgacgagccgggcgtgctgggccgcgacgaggcg 420
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 57041 tccgacacctacaccttctactcggccgacgagccgggcgtgctgggccgcgacgaggcg 56982
```

```
Query: 421   cgccccggcggggccgaggcggctccctccaggcgctggctggagcgcatgcggcggacc 480
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 56981 cgccccggcggggccgaggcggctccctccaggcgctggctggagcgcatgcggcggacc 56922


Query: 481   ttcgaggagcccacgtcgtcgctggccgcgcagatcctggctagcgtgtcggtggtgttc 540
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 56921 ttcgaggagcccacgtcgtcgctggccgcgcagatcctggctagcgtgtcggtggtgttc 56862


Query: 541   gtgatcgtgtccatggtggtgctgtgcgccagcacgttgcccgactggcgcaacgcagcc 600
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 56861 gtgatcgtgtccatggtggtgctgtgcgccagcacgttgcccgactggcgcaacgcagcc 56802


Query: 601   gccgacaaccgcagcctggatgaccggagcagg 633
              |||||||||||||||||||||||||||||||
Sbjct: 56801 gccgacaaccgcagcctggatgaccggagcagg 56769
```

☐ 1: AC025750. Reports Homo sapiens BAC ...[gi:18098549] Links

```
LOCUS       AC025750               151367 bp    DNA     linear   PRI 09-JAN-2002
DEFINITION  Homo sapiens BAC clone RP11-804P20 from 2, complete sequence.
ACCESSION   AC025750
VERSION     AC025750.10  GI:18098549
KEYWORDS    HTG.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Catarrhini;
            Hominidae; Homo.
REFERENCE   1  (bases 1 to 151367)
  AUTHORS   Sulston,J.E. and Waterston,R.
  TITLE     Toward a complete human genome sequence
  JOURNAL   Genome Res. 8 (11), 1097-1108 (1998)
   PUBMED   9847074
REFERENCE   2  (bases 1 to 151367)
  AUTHORS   Tomlinson,C., Cotton,M. and Doebber,A.
  TITLE     The sequence of Homo sapiens BAC clone RP11-804P20
  JOURNAL   Unpublished (2002)
REFERENCE   3  (bases 1 to 151367)
  AUTHORS   Waterston,R.H.
  TITLE     Direct Submission
  JOURNAL   Submitted (13-MAR-2000) Genome Sequencing Center, Washington
            University School of Medicine, 4444 Forest Park Parkway, St. Louis,
            MO 63108, USA
REFERENCE   4  (bases 1 to 151367)
  AUTHORS   Waterston,R.
  TITLE     Direct Submission
  JOURNAL   Submitted (09-JAN-2002) Department of Genetics, Washington
            University, 4444 Forest Park Avenue, St. Louis, Missouri 63108, USA
COMMENT     On Jan 9, 2002 this sequence version replaced gi:13431263.
            -------------- Genome Center
                Center: Washington University Genome Sequencing Center
                Center code: WUGSC
                Web site: http://genome.wustl.edu/gsc
                Contact: sapiens@watson.wustl.edu
            -------------- Summary Statistics
                Center project name: H_NH0804P20
            --------------.

            NOTICE:  This sequence may not represent the entire insert of this
            clone.  It may be shorter because we only sequence overlapping
            clone sections once, or longer because we provide a small overlap
            between neighboring data submissions.
```

EXHIBIT Z

Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities

PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
LinkOut
My NCBI (Cubby)

Related Resources
Order Documents
NLM Catalog
NLM Gateway
TOXNET
Consumer Health
Clinical Alerts
ClinicalTrials.gov
PubMed Central

1: Dev Biol. 1988 Nov;130(1):285-93.

Related Articles, Links

Regulated expression of a vitellogenin fusion gene in transgenic nematodes.

Spieth J, MacMorris M, Broverman S, Greenspoon S, Blumenthal T.

Program in Molecular, Cellular, and Developmental Biology, Indiana University, Bloomington 47405.

In Caenorhabditis elegans the vitellogenin genes are expressed abundantly in the adult hermaphrodite intestine, but are otherwise silent. In order to begin to understand the mechanisms by which this developmental regulation occurs, we used the transformation procedure developed for C. elegans by A. Fire (EMBO. J., 1986, 5, 2673-2680) to obtain regulated expression of an introduced vitellogenin fusion gene. A plasmid with vit-2 upstream and coding sequences fused to coding and downstream sequences of vit-6 was injected into oocytes and stable transgenic strains were selected. We obtained seven independent strains, in which the plasmid DNA is integrated at a low copy number. All strains synthesize substantial amounts of a novel vitellogenin-like polypeptide of 155 kDa that accumulates in the intestine and pseudocoelom, but is not transported efficiently into oocytes. In two strains examined in detail the fusion gene is expressed with correct sex, tissue, and stage specificity. Thus we have demonstrated that the nematode transgenic system can give proper developmental expression of introduced genes and so can be used to identify DNA regulatory regions.

PMID: 3181632 [PubMed - indexed for MEDLINE]

Proc. Natl. Acad. Sci. USA
Vol. 77, No. 12, pp. 7380–7384, December 1980
Genetics

EXHIBIT AA

Genetic transformation of mouse embryos by microinjection of purified DNA

(gene transfer/mice)

JON W. GORDON*, GEORGE A. SCANGOS†, DIANE J. PLOTKIN*, JAMES A. BARBOSA‡, AND FRANK H. RUDDLE*‡

*Department of Biology and ‡Department of Human Genetics, Yale University, New Haven, Connecticut 06511

Contributed by Frank H. Ruddle, September 23, 1980

ABSTRACT A recombinant plasmid composed of segments of herpes simplex virus and simian virus 40 viral DNA inserted into the bacterial plasmid pBR322 was microinjected into pronuclei of fertilized mouse oocytes. The embryos were implanted in the oviducts of pseudopregnant females and allowed to develop to term. DNA from newborn mice was evaluated by the Southern blotting technique for the presence of DNA homologous to the injected plasmid. Two of 78 mice in one series of injections showed clear homology, though the injected sequences had been rearranged. Band intensities from the two positive mice were consistent with the presence of donor DNA in most or all of the cells of the newborns. These results demonstrate that genes can be introduced into the mouse genome by direct insertion into the nuclei of early embryos. This technique affords the opportunity to study problems of gene regulation and cell differentiation in a mammalian system by application of recombinant DNA technology.

Introduction of specified gene sequences into mammalian embryos can be a powerful tool for the study of developmental genetic problems. The fate of such genes can be monitored throughout development by using sensitive probing techniques offered by recombinant DNA technology. In addition, the functioning of foreign genes in a normal host environment can be used to study the processes of gene regulation and to study the physiologic roles of products of such genes more precisely. Introduction of foreign DNA into all cells of an intact animal also provides an opportunity to pass sequences to offspring and to generate large numbers of transformed animals. In order to realize these benefits, it is necessary to transform embryos early in development and allow integration of foreign DNA into the cellular progenitors of the entire animal.

Such experiments with mammals are difficult. Zygotes must be maintained in culture conditions that at least grossly approximate the oviductal environment. Moreover, they can be maintained *in vitro* for only a few days, after which they must be returned to a female for implantation and further development. Insertion of material into early mammalian embryos is also difficult because of their small size.

Investigators have recently succeeded in constructing mosaic mice composed in part of descendants from cultured teratocarcinoma cells (1–3). This advance makes possible the introduction of genes into cultured cells, which might then be induced to cooperate in the formation of an intact adult mouse (4, 5). These cultured cells are often aneuploid, however, and some difficulty has been encountered in obtaining functional germ cells derived from them (6). Another problem with teratoma mosaics is that they are, indeed, mosaics. Thus, teratoma cells of XX chromosomal constitution cannot make sperm in mice that develop as males; the possibility of germ-line transmission in this system is accordingly reduced. Jaenisch and Mintz (7) have provided evidence that whole DNA of simian virus 40 (SV40), when placed in cavities of mouse blastocysts, may be found in the resultant offspring. Ideally, however, one would like to introduce a small amount of well-defined genetic material directly into normal embryos and allow this material to integrate and function within the host genome.

We have approached this problem by injecting DNA directly into the pronuclei of fertilized mouse oocytes. The one-cell stage was chosen in order to limit as much as possible the development of mosaicism during cleavage. To avoid the hazards of culture, injected embryos were immediately implanted into the oviducts of pseudopregnant recipients. The DNA chosen for injection was the bacterial plasmid pBR322 into which had been inserted fragments of herpes simplex and SV40 viral DNA. This plasmid was constructed because the SV40 fragment is known to contain an origin of DNA replication, whereas the herpes fragment codes for a gene product, thymidine kinase (TK), distinguishable from the endogenous mouse enzyme. DNA was extracted from newborn mice and screened by the Southern blotting technique for the presence of sequences homologous to the injected plasmid. Two of 78 mice evaluated in one experimental series were found to contain such sequences. In both instances the injected DNA had been modified, but it could be demonstrated to be derived from donor material. The intensity of the positive bands indicated that an amount of DNA roughly equivalent to one copy in every cell of the newborns was retained. We thus provide evidence that mice can be genetically transformed by direct insertion of DNA into early embryos.

MATERIALS AND METHODS

Mice. CD-1 mice were obtained from the Charles River Breeding Laboratories. B6D2F$_1$ mice were obtained from the Jackson Laboratory. All mice were maintained on a 14:10 light–dark schedule (lights off at 10 p.m., on at 8 a.m.). Six-week-old females were induced to superovulate with 5 international units of pregnant mares' serum (Gestyl, Organon) at 4 p.m. followed 48 hr later by 2.5 international units of human chorionic gonadotropin (Pregnyl, Organon) and placed immediately with males for mating. B6D2F$_1$ females were mated with CD-1 males; CD-1 females were mated with B6D2F$_1$ males. On the same evening other mature CD-1 females were placed with vasectomized CD-1 males. On the morning after mating (day 0) all female mice were examined for vaginal plugs. Six-week-old females were killed at 2 p.m. on day 0 and

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

Abbreviations: SV40, simian virus 40; TK, thymidine kinase; kb, kilobase(s).

† Present address: Department of Biology, The Johns Hopkins University, Baltimore, MD 21218.

their oviducts were removed into Krebs–Ringer bicarbonate-buffered medium supplemented with bovine serum albumin (8) and hyaluronidase at 1 mg/ml. Oviducts were opened with forceps and the fertilized eggs with remaining follicle cells were expressed into the dish. After 1–2 min, eggs were removed and washed three times in 2 ml of culture medium equilibrated with 5% CO_2 in air at 37°C. Eggs containing pronuclei were identified under the dissecting microscope and placed in lots of 20 in a microdrop of equilibrated medium, which was placed in a 100-mm tissue culture dish and covered with mineral oil (Mallinckrodt 6358). Eggs were stored in this manner in the incubator until microinjected.

Microinjection. Microneedles were pulled from thin-walled no. 1211L Omega Dot tubing (Glass Co. of America) on a DK1 model 700C pipette puller. Holding pipettes were pulled by hand on a microburner from G-12 capillary tubing (Thomas), and fire polished on a Sensaur microforge. The tips of the microneedles were allowed to fill with plasmid suspension by capillary action and the barrels were then filled with Fluorinert (3M FC77). They were then secured in PE-190 intramedic tubing on a Leitz micromanipulator. Holding pipettes were also filled with Fluorinert and similarly secured in PE-90 tubing. The tubing was likewise filled with Fluorinert and attached to 1-cm^3 Hamilton syringes. All manipulations were carried out on a Leitz microscope.

Tissue culture dishes containing the fertilized eggs were placed on the microscope and eggs were positioned by holding the pipette such that a pronucleus near the plasma membrane was close to the microneedle. The microneedle was inserted into the pronucleus and enough plasmid suspension was injected to cause an approximate doubling of the pronuclear volume (approximately 1 pl). Eggs that survived microinjection were removed and stored in a 30-mm tissue culture dish containing 2 ml of equilibrated medium until all microinjections were completed. Injection of 40–60 embryos required 1–2 hr.

Implantation. Plugged pseudopregnant CD-1 females were anesthetized with Nembutal at 6 mg/100 g of body weight. Ovaries were located through a dorsal incision. The ovarian bursa was torn away with no. 5 Dumont watchmaker's forceps, taking care not to rupture large blood vessels. The ostium of the oviduct was visualized under the dissecting microscope and a pipette containing 10–20 microinjected embryos was inserted into it. The eggs were expelled into the oviduct and the wound was closed with wound clips. Mice were examined on days 18–21 for the delivery of live offspring. Newborn mice were stored at −80°C for later analysis. Sixty percent of the embryos survived microinjection; 30–50% of the survivors developed into live young. All newborns were normal in appearance. All microinjection work was carried out under P1 containment in accordance with National Institutes of Health guidelines.

DNA Isolation. DNA was isolated from whole newborn mice by the method of Blin and Stafford (9) with the following modifications. Powdered tissue was incubated for 4 hr at 50°C in 22 ml of 0.28 M EDTA/0.5% Sarkosyl, pH 7.0. The homogenate was subsequently extracted twice in phenol/chloroform/isoamyl alcohol (15 ml:5 ml:0.2 ml), and once in chloroform/isoamyl alcohol (15 ml:0.6 ml). The extract was dialyzed for 24 hr against 10 mM Tris·HCl, pH 8.0/10 mM NaCl/1 mM EDTA and precipitated with a 2-vol excess of 100% ethyl alcohol. Precipitated DNA was stored at −20°C until use.

Filter Hybridization. DNA was redissolved in 1× TEN (10 mM Tris·HCl, pH 7.75/10 mM NaCl/0.1 mM EDTA) to yield a final concentration of approximately 1 mg/ml. Twenty micrograms of DNA was digested at a 10- to 20-fold excess with appropriate restriction enzymes (Bethesda Research Laboratories, Rockville, MD). After overnight digestion at 37°C, samples were electrophoresed in 1% agarose in 160 mM Tris·HCl/80 mM NaOAc/80 mM NaCl/5 mM EDTA, pH 8, at 350 A for 22 hr. Samples were then blotted onto nitrocellulose filters according to the method of Southern (10).

Nick translations were performed by using the New England Nuclear nick translation kit with ^{32}P-labeled dCTP obtained from New England Nuclear. Filter hybridizations were performed as described by Wahl *et al.* (11). Filters were then used to expose Kodak X-Omat x-ray film, using intensifying screens, until band intensities were appropriate for analysis.

Construction of the Plasmid. The recombinant plasmids, called pST6, pST9 and pST12, carrying the SV40 origin of replication and promoters, and the herpes simplex virus TK gene were constructed by inserting the SV40 *Hin*dIII-C fragment (12, 13) into the available *Hin*dIII site in the plasmid pTKX-1 (14). DNA from the SV40 mutant 1265, kindly provided by C. Cole of Yale University, was digested to completion with restriction enzymes *Hin*dIII and *Hin*fI (New England BioLabs) simultaneously. The double digestion generated two fragments larger than 550 base pairs; the *Hin*dIII-C fragment (1099 base pairs; map position 0.649–0.859) and the *Hin*fI-B fragment (1085 base pairs; 0.992–0.199), which comigrated on a 1% Seaplaque agarose gel. The 1.1-kilobase (kb) doublet band was extracted from the gel and ligated with pTKX-1 that had been digested with *Hin*dIII and alkaline phosphatase [as described by Ullrich *et al.* (15) except that bovine alkaline phosphatase (Sigma) was used]. The molar ratio of the vector to target in the ligation mixture was 3:1. The ligation mixture was incubated at 4°C for 17 hr with one addition of phage T4 ligase at 11 hr. The mixture was used to transform *Escherichia coli* strain HB 101, and ampicillin-resistant colonies were selected. Colonies carrying the putative pST plasmids were identified by colony hybridization (16), using SV40 DNA as the probe. Approximately 20% of the ampicillin-resistant colonies contained SV40 sequences. Confirmation of the *Hin*dIII-C fragment insertion and determination of its orientation in the plasmid was done by restriction analysis of mini-DNA isolations (17). A restriction endonuclease map of the plasmid pST6 is shown in Fig. 1. This work was carried out under P2 containment in accordance with National Institutes of Health guidelines.

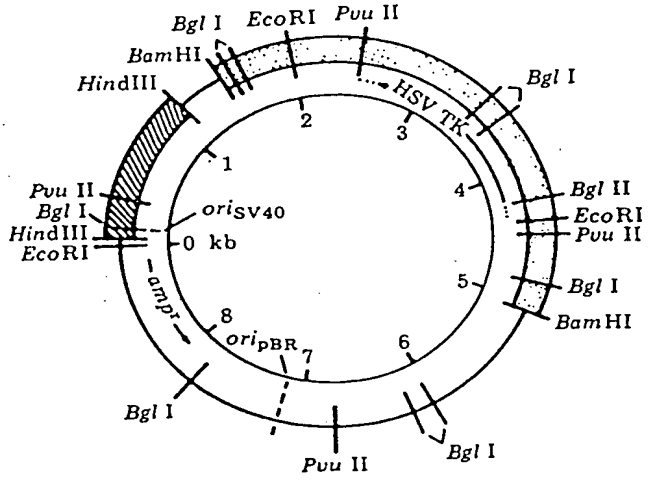


FIG. 1. The circular plasmid pST6, a derivative of pBR322. Hatched area shows the SV40 insert; stippled area denotes the herpes simplex virus TK insert. *amp*^{r}, ampicillin resistance gene; *ori*, origin of DNA replication.

Table 1. Summary of microinjection data

Exp.	Plasmid	Copies injected per cell	Offspring	Plasmid DNA positives
1	pST6	1,000	78	2
2	pST6	12,000	10	0
3	pST6 (linearized)	1,000	40	0
4	pST9	1,000	16	0
5	pRH 1.3Mm 1	1,000	27	0
6	pST12	500	2	0
7	Uninjected control	—	54	0

The pRH 1.3Mm 1 plasmid consists of a cloned fragment of a member of the highly repeated and interspersed *Eco*RI–*Bgl* II sequence family cloned in pBR322, provided by N. Arnheim (18). pST9 is identical to pST6, except that the orientation of the SV40 fragment is reversed. pST12 is a dimer of pST6. pST was linearized by *Sal* I digestion. A total of 187 mice were born from microinjected embryos.

RESULTS

Results of the plasmid microinjections are summarized in Table 1. In the first experimental series, injection of several hundred embryos yielded 78 live young. DNA was extracted from whole newborn mice for rapid and efficient determination of transformation frequency. The screening method gives a low estimate of the number of transformants; embryos with transforming DNA in a small percentage of their cells could have escaped detection. DNA from 2 of these 78 newborn mice contained sequences that hybridized strongly with the probe, pST6. The restriction endonuclease patterns of the incorporated sequences were significantly different between the two offspring, and are described below.

DNA from the first positive animal, no. 48, gave two intense bands with estimated sizes of 12.9 kb and 9.8 kb and a third band of very large size (>24 kb) when digested with *Bam*HI (Fig. 2). The positions of the two smaller bands were unaffected by digestion with *Hin*dIII, *Eco*RI, *Bam*HI, or *Xba* I (Fig. 2). This result suggested that the TK sequences, which had been inserted into the *Bam*HI sites, and the SV40 sequences inserted into the *Hin*dIII sites were not present in their native state in the incorporated material. The *Hin*dIII digestion, however, was incomplete as judged from the control track. We therefore probed with SV40 DNA alone. No sequences homologous to this

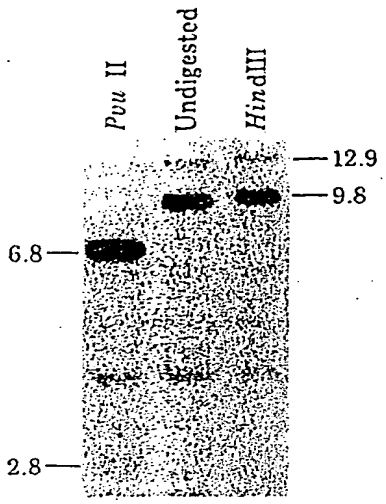


FIG. 3. DNA from mouse no. 48 digested with *Hin*dIII or *Pvu* II, or undigested; probed with pST6. Fragment sizes are indicated in kb.

probe were detected. The 12.9- and 9.8-kb fragments appeared in the undigested sample, consistent with their presence as free molecules. Digestion of the DNA with *Pvu* II generated two bands of altered mobility, 2.8 kb and 6.8 kb in size (Fig. 3). This result indicated that the sequences represented by the 12.9- and 9.8-kb bands contained at least one *Pvu* II site. We believe these results, taken together, are consistent with the existence of free circular molecules in the DNA of mouse no. 48.

The second positive, no. 73, showed a markedly different blotting pattern. In the undigested DNA, hybridizable material was not separable from the high molecular weight mouse DNA. Moreover, digestion with *Xba* I, which does not cut pST6, gave a single band of greater size than the highest molecular weight standard of 23.7 kb. Finally, several bands showed homology with probes synthesized from either purified SV40 DNA or TK fragment (Fig. 4). Thus, this animal had retained all or part of these portions of the plasmid.

Digestion with *Bam*HI yielded three major bands, 7.8 kb, 3.9 kb, and 3.4 kb. The largest band showed homology with

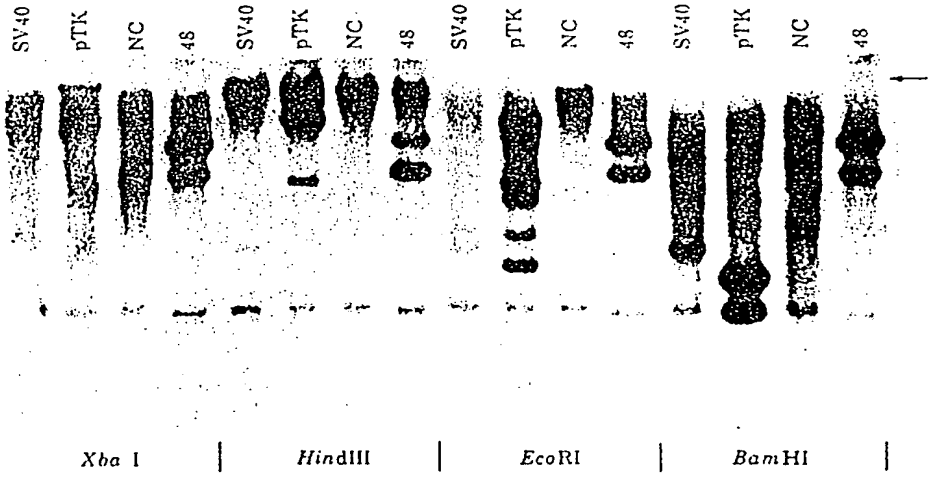


FIG. 2. DNA from mouse no. 48 digested with *Bam*HI, *Eco*RI, *Hin*dIII, and *Xba* I. The labeled probe was pST6 DNA. NC indicates the negative control (DNA isolated from uninjected mice). Positive controls include (*i*) NC DNA with SV40 DNA added (SV40) and (*ii*) NC DNA with the plasmid pTTX-1 added (pTK). Arrow indicates the high molecular weight band that appears reproducibly in *Bam*HI digests.

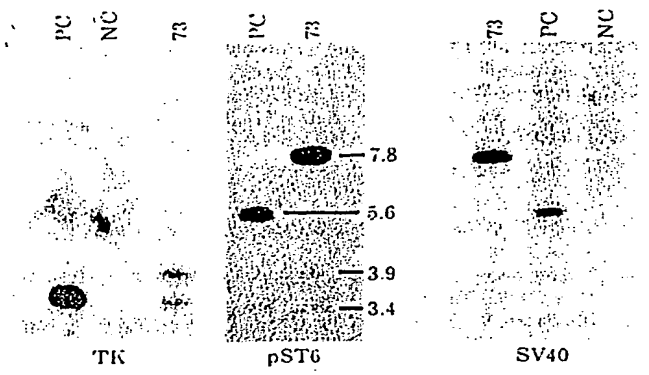


FIG. 4. DNA from mouse no. 73 digested with *Bam*HI and probed with pST6 (*Center*), SV40 DNA (*Right*), or TK fragment (*Left*). Positive control (PC) consists of pST6 added to mouse DNA to a concentration of 10^{-6} by weight. NC denotes the negative control (DNA isolated from uninjected mice).

SV40, SV40 + pBR322, and with the whole plasmid, pST6 (Fig. 4). The two smaller bands showed homology with TK fragment, but not with SV40 or pBR322 (Fig. 4). The probeable portions of these smaller pieces were thus composed entirely of TK-derived material. The smallest band, 3.4 kb, closely approximates the size of the TK fragment that was inserted into the *Bam*HI sites, suggesting that the entire TK gene had been retained in no. 73. Digestion with *Bgl* I, however, proved this supposition incorrect. An internal fragment of TK approximately 1.8 kb in size, defined by *Bgl* I sites, did not appear in the DNA (data not shown). This showed that the 3.4- and 3.9-kb *Bam*HI fragments were composed of portions of the TK fragment that had either been concatamerized or complexed with mouse DNA to yield molecular weights equal to or greater than the molecular weight of the original TK insert.

Digests with *Pvu* II and *Hin*dIII provided strong evidence that the entire SV40 sequence was retained. Digestion with *Hin*dIII produced a fragment very close in size to the SV40 insert of pST6. In addition, digestion with *Pvu* II gave two fragments that migrated indistinguishably from the *Pvu* II-defined SV40 fragments of pST6. Thus, two independent experiments support the contention that the entire SV40 fragment was present.

DISCUSSION

These data demonstrate that it is possible to use a recombinant plasmid as a vector for transfer of foreign genes directly into mouse embryos, and that these embryos can maintain the foreign genes throughout development. Moreover, the intensity of the bands on Southern blot analysis suggests that most or all of the cells of the newborns contained derivatives of the injected plasmid. Blotting experiments with hybrid cell populations have shown that sequences cannot be detected if present in fewer than 10% of the cells (19). We are thus confident that the two transformed mice contained enough plasmid DNA for distribution of one copy to at least this percentage of their cells. Our positive controls were adjusted to correspond to one copy of pST6 per diploid genome. The band intensities of no. 48 and no. 73 are comparable to the control. Thus, the transforming sequences are probably present in far greater amounts than the 10% threshold of detectability; the band intensities are more consistent with the presence of the plasmid derivative in most or all of the cells of the newborns. Our method of analysis cannot rule out the possibility that only a few of the cells contained all of the sequences while most of the cells were negative, but we consider unlikely the chances that cells carrying a large amount of additional genetic material would survive and compete successfully through development. If the transforming sequences were in fact distributed throughout the tissues of the mice, then integration must have occurred at an early stage, shortly after determination of the inner cell mass. Injection of one-celled embryos may be important for obtaining early integration. In addition, the high mortality caused by microinjection suggests that injection of only a fraction of the cells of a later cleavage stage might result in preferential survival of uninjected blastomeres and consequently give a lower rate of success.

The transformation rate reported here compares very favorably with other gene transfer systems involving mammalian cells. Calcium phosphate-mediated gene transfer into cultured cells results in transformation rates of 10^{-8} to 10^{-5} (20, 21), while microinjection of cultured cells gives approximately 5% success (22). Our transformation rate agrees well with these latter results. The reasons for higher rates in microinjection experiments are unknown but may include the facts that DNA is inserted directly into the nucleus and that gene expression is not required in the mouse system.

Significant differences were found between the two transformed mice. In mouse no. 48, SV40 and herpes viral TK DNA could not be detected. The remaining sequences, derived from pBR322, were complexed into three bands, all of higher molecular weight than the entire pBR322 plasmid. In addition, two of these bands represented DNA that probably existed free of the host genome. The presence of unintegrated sequences in no. 48 is intriguing. Two plausible models can be invoked to explain this observation: (*i*) these sequences may have replicated autonomously and persisted as plasmid-like units; (*ii*) alternatively, they may have been generated from an integrated segment. The former model requires that the free sequences have the capacity to replicate. The plasmid from which they descended did contain the pBR322 and SV40 origins. But, interestingly, SV40 DNA is undetectable in the retained material. It is also possible that a mouse origin was acquired as a result of interaction with the host genome.

It is more likely that the free sequences were generated from integrated material. Generation of free circular DNA from transformed cultured cells has been observed previously (23). Cells infected with viruses can also generate free DNA from the integrated viral genome (24). In addition, cells transformed in calcium phosphate-mediated gene transfer experiments can pass through an unstable phase during which the donated material is maintained independent of the host genome as high molecular weight "transgenomes" (25). An important characteristic of these independent transgenomes is their rapid loss from recipient cells; as many as 10% of the cells may lose the transforming sequences per day (25). The rearrangement of the donor material in no. 48 appears analogous to transgenome formation in cultured cells. If the unintegrated sequences were similar to independent transgenomes, we would expect them to be rapidly lost from the mouse cells during development and not detectable in the newborn. The marked intensity of the two bands in no. 48 rather suggests that they were continuously being produced from an integrated sequence. The presence of a high molecular weight band after digestion with *Bam*HI is also consistent with the integration model. This band may represent material from which the two smaller bands were generated.

In mouse no. 73, no free sequences were present. Both the undigested and *Xba* I-digested samples gave single bands of greater size than the highest molecular weight standard. Moreover, SV40 and TK sequences were retained in this animal. The patterns of bands present in mouse no. 73 is explained best

by plasmid integration into the mouse genome at a site within the TK region. In this model, digestion of the mouse DNA with *Bam*HI would generate three plasmid-derived fragments, two of which would consist of the TK fragment (now at both ends of the integrated molecule) linked to mouse DNA. The third fragment would be cleaved from within the integrated plasmid and would contain the SV40 and pBR322 moieties. The predicted size of this internal fragment is 5.5 kb. This model also predicts that the TK fragment would be disrupted and that the SV40 and pBR322 sequences would be intact. The DNA of mouse no. 73 contained two bands of 3.4 and 3.9 kb that hybridized only with the purified TK fragment and contained no sequences homologous to SV40 or pBR322, and a band of 7.8 kb that hybridized to SV40 and not to TK. The large size of this fragment relative to the expected 5.5-kb fragment might be due to partial internal duplication, which is consistent with independent observations of SV40 integration (26, 27). Digestion of the DNA of mouse no. 73 with *Bgl* I or with *Pvu* II failed to generate expected fragments from within the TK insert but indicated that most or all of pBR322 and SV40 were present. Additionally, *Hin*dIII digestion generated a band of the expected size of the SV40 insert, indicating that all of the SV40 sequences present on pST6 were also present in the DNA of mouse no. 73 (data not shown). Thus, our observations are consistent with a single integration event.

An important similarity between the two positive mice was the extensive rearrangement of the sequences. In the first instance, SV40 and herpes virus TK sequences were largely if not entirely removed from the injected DNA. In the second case, SV40 sequences and herpes virus TK sequences were demonstrable, but the TK gene was significantly rearranged. These observations raise the possibility that selection occurred against embryos that retained the TK gene intact and in an active state. The possibility that herpes virus TK is teratogenic to mouse embryos is consistent with our data. We consider this notion unlikely, however, because cells transformed in culture and under selection for TK demonstrate similar patterns of rearrangement (25, 28).

These initial results show that genetic transformation can be extended to whole mammalian organisms at a very early stage in their development. Further refinement of these techniques should lead to a reliable system of embryo transformation with its attendant applications for investigation of problems in development and cell differentiation.

Note Added in Proof. We have produced a third transformant by injection of 30,000 copies per cell of the plasmid pST9. Restriction analysis indicates that, as in mouse no. 73, the transforming sequences are integrated. Initial studies also indicate that at least one complete copy each of both the herpes virus TK and SV40 regions has been retained in this animal.

We thank Dr. N. Arnheim, of the State University of New York at Stony Brook, for supplying us with the pRH 1.3Mm 1 plasmid, B. Kay for advice on construction of pST plasmids, K. M. Huttner for helpful discussions, S. Pafka for photography, and M. Siniscalchi for typing this manuscript. This work was supported by National Institutes of Health Grant GM09966-19 to F.H.R., and G.A.S. was supported by National Institutes of Health Fellowship GM06528-01. J.W.G. was supported during 1979–1980 by a Hudson Brown Fellowship administered through the Department of Obstetrics–Gynecology, Yale University School of Medicine, and is currently supported by National Institutes of Health Grant IF32GM07959-01. D.J.P. was supported by Damon Runyon–Walter Winchell Cancer Fund Grant DRG 291 F and is currently supported by U.S. Public Health Service Grant 3F32 CA06414-0151. J.A.B. is supported by National Institutes of Health Training Grant 5T32-HD07149-03.

1. Brinster, R. L. (1974) *J. Exp. Med.* 140, 1049–1056.
2. Mintz, B. & Illmensee, K. (1975) *Proc. Natl. Acad. Sci. USA* 72, 3585–3589.
3. Papaioannou, V. E., McBurney, M. W., Gardner, R. L. & Evans, M. S. (1975) *Nature (London)* 258, 69–73.
4. Illmensee, K., Hoppe, P. C. & Croce, C. M. (1978) *Proc. Natl. Acad. Sci. USA* 75, 1914–1918.
5. Pellicer, A., Wagner, E. F., Kareh, A. E., Dewey, M. J., Reuser, A. J., Silverstein, S., Axel, R. & Mintz, B. (1980) *Proc. Natl. Acad. Sci. USA* 77, 2098–2102.
6. Illmensee, K. (1978) in *Genetic Mosaics and Chimeras in Mammals*, ed., Russell, L. B. (Plenum, New York), pp. 3–25.
7. Jaenisch, R. & Mintz, B. (1974) *Proc. Natl. Acad. Sci. USA* 71, 1250–1254.
8. Mullen, R. J. & Whitten, W. K. (1971) *J. Exp. Zool.* 178, 165–176.
9. Blin, N. & Stafford, D. W. (1976) *Nucleic Acids Res.* 3, 2303–2308.
10. Southern, E. M. (1975) *J. Mol. Biol.* 98, 503–517.
11. Wahl, G. M., Stern, M. & Stark, G. R. (1979) *Proc. Natl. Acad. Sci. USA* 76, 3683–3687.
12. Reddy, V. B., Thimmappaya, B., Dahr, R., Subramanian, K. N., Zain, B. S., Pan, J., Ghosh, P. K., Celma, B. L. & Weissmann, S. W. (1978) *Science* 200, 494–502.
13. Fiers, W., Contreras, R., Haegeman, G., Rogiers, R., Van de Voorde, A., Van Heuverswyn, A., Van Herreweghe, J., Volckaert, G. & Ysebaert, M. (1978) *Nature (London)* 273, 113–120.
14. Enquist, L. W., Van de Woude, G. F., Wagner, M., Smiley, J. R. & Summers, W. C. (1979) *Gene* 7, 335–342.
15. Ullrich, A., Shine, J., Chirgwin, J., Pictet, R., Tischer, E., Rutter, W. J. & Goodman, H. M. (1977) *Science* 196, 1313–1319.
16. Gergen, J. P., Stern, R. H. & Wensink, P. C. (1979) *Nucleic Acids Res.* 7, 2115–2136.
17. Birnboim, H. C. & Doty, J. (1979) *Nucleic Acids Res.* 7, 1513–1523.
18. Heller, R. & Arnheim, N. (1980) *Nucleic Acids Res.*, in press.
19. D'Eustachio, P., Pravtcheva, D., Marcu, K. & Ruddle, F. H. (1980) *J. Exp. Med.* 151, 1545–1550.
20. Wigler, M., Pellicer, A., Silverstein, S. & Axel, R. (1978) *Cell* 14, 725–731.
21. Lester, S. C., LeVan, S. K., Steglich, C. & DeMars, R. (1980) *Somatic Cell Genet.* 6, 241–259.
22. Anderson, W. F., Killos, L., Sanders-Haigh, L., Kretchmer, P. J. & Diacumakos, E. G. (1980) *Proc. Natl. Acad. Sci. USA* 77, 5399–5403.
23. Schegget, J. T., Voves, J., Strien, A. V. & Van der Noordaa, J. (1980) *J. Virol.* 35, 331–339.
24. Hanahan, D., Lane, D., Lipsich, L., Wigler, M. & Botchan, M. (1980) *Cell* 21, 127–140.
25. Scangos, G. A., Huttner, K. M., Juricek, D. K. & Ruddle, F. H., *Mol. Cell. Biol.*, in press.
26. Kelly, T. J., Jr., Lewis, A. M., Jr., Levine, A. S. & Siegel, S. (1974) *J. Mol. Biol.* 89, 113–126.
27. Botchan, M., Topp, W. & Sambrook, J. (1979) *Cold Spring Harbor Symp. Quant. Biol.* 43, 709–719.
28. Huttner, K. M., Scangos, G. A. & Ruddle, F. H. (1979) *Proc. Natl. Acad. Sci. USA* 76, 5820–5824.

EXHIBIT BB

About Entrez
Text Version

Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities

PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
LinkOut
My NCBI (Cubby)

Related Resources
Order Documents
NLM Catalog
NLM Gateway
TOXNET
Consumer Health
Clinical Alerts
ClinicalTrials.gov
PubMed Central

1: Nature. 1990 Apr 5;344(6266):541-4. Related Articles, Links

Fulminant hypertension in transgenic rats harbouring the mouse Ren-2 gene.

Mullins JJ, Peters J, Ganten D.

German Institute for High Blood Pressure Research, University of Heidelberg.

PRIMARY hypertension is a polygenic condition in which blood pressure is enigmatically elevated; it remains a leading cause of cardiovascular disease and death due to cerebral haemorrhage, cardiac failure and kidney disease. The genes for several of the proteins involved in blood pressure homeostasis have been cloned and characterized, including those of the renin-angiotensin system, which plays a central part in blood pressure control. Here we describe the introduction of the mouse Ren-2 renin gene into the genome of the rat and demonstrate that expression of this gene causes severe hypertension. These transgenic animals represent a model for hypertension in which the genetic basis for the disease is known. Further, as the transgenic animals do not overexpress active renin in the kidney and have low levels of active renin in their plasma, they also provide a new model for low-renin hypertension.

PMID: 2181319 [PubMed - indexed for MEDLINE]

EXHIBIT CC

Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities

PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
LinkOut
My NCBI (Cubby)

Related Resources
Order Documents
NLM Catalog
NLM Gateway
TOXNET
Consumer Health
Clinical Alerts
ClinicalTrials.gov
PubMed Central

1: Nature. 1985 Jun 20-26;315(6021):680-3.

Related Articles, Links

Production of transgenic rabbits, sheep and pigs by microinjection.

Hammer RE, Pursel VG, Rexroad CE Jr, Wall RJ, Bolt DJ, Ebert KM, Palmiter RD, Brinster RL.

Direct microinjection has been used to introduce foreign DNA into a number of terminally differentiated cell types as well as embryos of several species including sea urchin, Candida elegans, Xenopus, Drosophila and mice. Various genes have been successfully introduced into mice including constructs consisting of the mouse metallothionein-I (MT) promoter/regulator region fused to either the rat or human growth hormone (hGH) structural genes. Transgenic mice harbouring such genes commonly exhibit high, metal-inducible levels of the fusion messenger RNA in several organs, substantial quantities of the foreign growth hormone in serum and enhanced growth. In addition, the gene is stably incorporated into the germ line, making the phenotype heritable. Because of the scientific importance and potential economic value of transgenic livestock containing foreign genes, we initiated studies on large animals by microinjecting the fusion gene, MT-hGH, into the pronuclei or nuclei of eggs from superovulated rabbits, sheep and pigs. We report here integration of the gene in all three species and expression of the gene in transgenic rabbits and pigs.

PMID: 3892305 [PubMed - indexed for MEDLINE]

RESEARCH PAPERS

Bio/Technology **8**, 140 - 143 (1990)
doi:10.1038/nbt0290-140

Rabbit β-Casein Promoter Directs Secretion of Human Interleukin-2 into the Milk of Transgenic Rabbits

Th. A. Bühler[1], Th. Bruyère[2], D. F. Went[1], G. Stranzinger[1] & K. Bürki[2]

[1]Swiss Federal Institute of Technology Zürich, Institute for Animal Science, CH-8092 Zürich.

[2]Preclinical Research, Sandoz Ltd., CH-4002 Basel.

To test the potential usefulness of transgenic rabbits as production systems for human proteins of pharmaceutical value, we cloned the rabbit β-casein promoter and fused it to the genomic sequence of the human interleukin-2 (hIL2) gene. Four transgenic female rabbits were tested for expression and biological activity of the foreign protein in their milk. The milk of all four females proved to contain biologically active hIL2. The results show that transgenic rabbits may represent a convenient and economic system for the rapid production of biologically active protein in milk.

REFERENCES

1. Gordon, K., Lee, E., Vitale, J.A., Smith, A.E., Westphal, H. and Hennighausen, L. 1987. Production of human tissue plasminogen activator in transgenic mouse milk. *Bio/Technology* **5**: 1183–1187. | ISI | ChemPort |
2. Pittius, C.W., Hennighausen, L., Lee, E., Westphal, H.,

inducible genes. *DNA* **5**: 383–391. | PubMed | ISI | ChemPort |
18. Brinster, R.L., Chen, H.Y., Trumbauer, M.E., Yagle, M.K. and Palmiter, R.D. 1985. Factors affecting the efficiency of introducing foreign DNA into mice by microinjecting eggs. *Proc. Natl. Acad. Sci. U.S.A.* **82**: 4438–4442. | PubMed | ChemPort |
19. Hammer, R.E., Pursel, V.G., Rexroad, C.E., Wall, R.J., Bolt, D.J., Ebert, K.M., Palmiter, R.D. and Brinster, R.L. 1985. Production of transgenic rabbits, sheep and pigs by microinjection. *Nature* **315**: 680–683. | PubMed | ISI | ChemPort |
20. Robb, R.J., 1982. Human T-cell growth factor: Purification, biochemical characterisation, and interaction with a cellular receptor. *Immunobiol.* **161**: 21–50. | ISI | ChemPort |
21. Lebas, F. 1970. Description d'une machine à traire les lapins. *Ann. Zootech.* **19**: 223–228. | ISI |
22. Lee, K.F., Atiee, S.H. and Rosen, J.M. 1989. Differential regulation of rat β-casem-chloraraphenicol acetyltransferase fusion gene expression in transgenic mice. *Mol. Cell. Biol.* **9**: 560–565. | PubMed | ISI | ChemPort |
23. Bornstein, P., McKay, J., Liska, D.J., Apone, S. and Devarayalu, S. 1988. Interactions between the promoter and the first intron are involved in transcriptional control of α1(I) collagen gene expression. *Mol. Cell. Biol.* **8**: 4851–4857. | PubMed | ISI | ChemPort |
24. Lee, K.F., DeMayo, F.J., Atiee, S.H. and Rosen, J.M. 1988. Tissue specific expression of the rat β-casein gene in transgenic mice. *Nucl. Acids Res.* **16**: 1027–1041. | PubMed | ISI | ChemPort |
25. Grosveld, F., van Assendelft, G.B., Greaver, D.R. and Kollias, G. 1987. Position-independent, high-level expression of the human β-globin gene in transgenic mice. *Cell* **51**: 975–985. | Article | PubMed | ISI | ChemPort |
26. Talbot, D., Collis, P., Antoniou, M., Vidal, M., Grosveld, F. and Greaves, D.R. 1989. A dominant control region from the human β-globin locus conferring integration site-independent gene expression. *Nature* **338**: 352–355. | Article | PubMed | ISI | ChemPort |
27. Zoller, M.J. and Smith, M. 1983. Oligonucleotide-directed mutagenesis of DNA fragments cloned into M13 vectors. *Methods in Enzymol.* **100**: 468–500. | Article | ISI | ChemPort |
28. Ogden, R.C. and Adams, D.A. 1987. Electrophoresis in agarose and acrylamide gels. *Methods in Enzymol.* **152**: 61–87. | Article | ISI | ChemPort |
29. Hogan, B., Costantini, F. and Lacy, E. 1986. *Manipulating the Mouse Embryo: A Laboratory Manual*. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
30. Schreier, M.H. and Tees, R. Long-term culture and cloning of specific helper T cells, p. 263–275. *In*; *Immunological Methods*. Vol. **2**. Academic Press, N.Y.

nature biotechnology

ISSN: 1087-0156
EISSN: 1546-1696

EXHIBIT EE

About Entrez
Text Version

Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities

PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
LinkOut
My NCBI (Cubby)

Related Resources
Order Documents
NLM Catalog
NLM Gateway
TOXNET
Consumer Health
Clinical Alerts
ClinicalTrials.gov
PubMed Central

1: Gene Ther. 2000 Jun;7(12):1046-54.

Related Articles, Links

nature publishing group

Effect of transgenic GDNF expression on gentamicin-induced cochlear and vestibular toxicity.

Suzuki M, Yagi M, Brown JN, Miller AL, Miller JM, Raphael Y.

Kresge Hearing Research Institute, The University of Michigan, Ann Arbor 48109-0648, USA.

Gentamicin administration often results in cochlear and/or vestibular hair cell loss and hearing and balance impairment. It has been demonstrated that adenovirus-mediated overexpression of glial cell line-derived neurotrophic factor (GDNF) can protect cochlear hair cells against ototoxic injury. In this study, we evaluated the protective effects of adenovirus-mediated overexpression of GDNF against gentamicin ototoxicity. An adenovirus vector expressing the human GDNF gene (Ad.GDNF) was administered into the scala vestibuli as a rescue agent at the same time as gentamicin, or as a protective agent, 7 days before gentamicin administration. Animals in the Rescue group displayed hearing thresholds that were significantly better than those measured in the Gentamicin or Ad.LacZ/Gentamicin groups. In the Protection group, Ad.GDNF afforded significant preservation of utricular hair cells. The data demonstrated protection of the inner ear structure, and rescue of the inner ear structure and function against ototoxic insults. These experiments suggest that inner ear gene therapy may be developed as a clinical tool for protecting the ear against environmentally induced insults.

PMID: 10871754 [PubMed - indexed for MEDLINE]

Article

JARO - Journal of the Association for Research in Otolaryngology
Publisher: Springer-Verlag New York, LLC
ISSN: 1525-3961 (Paper) 1438-7573 (Online)
DOI: 10.1007/s101620010011
Issue: Volume 1, Number 4
Date: December 2000
Pages: 315 - 325

Spiral Ganglion Neurons Are Protected from Degeneration by GDNF Gene Therapy

Masao Yagi, Sho Kanzaki, Kohei Kawamoto, Brian Shin, Pratik P. Shah, Ella Magal, Jackie Sheng, Yehoash Raphael

[A1] Kresge Hearing Research Institute, University of Michigan, Ann Arbor, MI 48109-0506, USA
[A2] Department of Neuroscience, Amgen Inc., Thousand Oaks, CA 91320, USA
[A3] Department of Otolaryngology, Kansai Medical University, Moriguchi, Osaka 570-8506, Japan
[A4] Department of Otolaryngology, Keio University, Shinjuku-ku, Tokyo 160-0016, Japan

Abstract:

Perceptual benefits from the cochlear prosthesis are related to the quantity and quality of the patient's auditory nerve population. Multiple neurotrophic factors, such as glial cell line-derived neurotrophic factor (GDNF), have been shown to have important roles in the survival of inner ear auditory neurons, including protection of deafferented spiral ganglion cells (SGCs). In this study, GDNF gene therapy was tested for its ability to enhance survival of SGCs after aminoglycoside/diuretic-induced insult that eliminated the inner hair cells. The GDNF transgene was delivered by adenoviral vectors. Similar vectors with a reporter gene (lacZ) insert served as controls. Four or seven days after bilateral deafening, 5 ml of an adenoviral suspension (Ad-GDNF or Ad-lacZ) or an artificial perilymph was injected into the left scala tympani of guinea pigs. Animals were sacrificed 28 days after deafening and their inner ears prepared for SGC counts. Adenoviral-mediated GDNF transgene expression enhanced SGC survival in the left (viral-treated) deafened ears. This observation suggests that GDNF is one of the survival factors in the inner ear and may help maintain the auditory neurons after insult. Application of GDNF and other survival factors via gene therapy has great potential for inducing survival of auditory neurons following hair cell loss.

EXHIBIT FF

Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities

PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
LinkOut
My NCBI (Cubby)

Related Resources
Order Documents
NLM Catalog
NLM Gateway
TOXNET
Consumer Health
Clinical Alerts
ClinicalTrials.gov
PubMed Central

1: Biochim Biophys Acta. 2000 Oct 18;1523(2-3):161-71. Related Articles, Links

ELSEVIER SCIENCE FULL-TEXT ARTICLE

Introduction of the human growth hormone gene into the guinea pig mammary gland by in vivo transfection promotes sustained expression of human growth hormone in the milk throughout lactation.

Hens JR, Amstutz MD, Schanbacher FL, Mather IH.

Department of Animal and Avian Sciences, University of Maryland, College Park 20742, USA.

We tested the feasibility of transfecting mammary tissue in vivo with an expression plasmid encoding the human growth hormone (hGH) gene, under the control of the cytomegalovirus promoter. Guinea pig mammary glands were transfected with plasmid DNA infused through the nipple canal and expression was monitored in control and transfected glands by radioimmunoassay of milk samples for hGH. Sustained expression of hGH throughout lactation was attained with a polyion transfection complex shown to be optimal for the transfection of bovine mammary cells, in vitro. However, contrary to expectations, hGH expression was consistently 5- to 10-fold higher when DEAE-dextran was used alone for transfection. Thus polyion complexes which are optimal for the transfection of cells in vitro may not be optimal in vivo. The highest concentrations of hGH in milk were obtained when glands were transfected within 3 days before parturition. This method may have application for studying the biological role or physical properties of recombinant proteins expressed in low quantities, or for investigating the regulation of gene promoters without the need to construct viral vectors or produce transgenic animals.

PMID: 11042380 [PubMed - indexed for MEDLINE]

EXHIBIT GG

Proc. Natl. Acad. Sci. USA
Vol. 88, pp. 1696–1700, March 1991
Agricultural Sciences

High-level synthesis of a heterologous milk protein in the mammary glands of transgenic swine

ROBERT J. WALL*, VERNON G. PURSEL*, AVI SHAMAY†, ROBERT A. MCKNIGHT†, CHRISTOPH W. PITTIUS†‡, AND LOTHAR HENNIGHAUSEN†§

*Reproduction Laboratory, Agricultural Research Service, U.S. Department of Agriculture, Beltsville, MD 20705; and †Laboratory of Biochemistry and Metabolism, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892

Communicated by Gary Felsenfeld, November 19, 1990 (received for review August 27, 1990)

ABSTRACT The whey acidic protein (WAP) is a major milk protein in mice, rats, and rabbits but has not been found in milk of livestock including swine. To determine whether mammary gland regulatory elements from the WAP gene function across species boundaries and whether it is possible to qualitatively alter milk protein composition, we introduced the mouse WAP gene into the genome of swine. Three lines of transgenic swine were analyzed, and mouse WAP was detected in milk from all lactating females at concentrations of about 1 g/liter; these levels are similar to those found in mouse milk. Expression of the corresponding RNA was specific to the mammary gland. Our results suggest that the molecular basis of mammary-specific gene expression is conserved between swine and mouse. In addition the WAP gene must share, with other milk protein genes, elements that target gene expression to the mammary gland. Mouse WAP accounted for about 3% of the total milk proteins in transgenic pigs, thus demonstrating that it is possible to produce high levels of a foreign protein in milk of farm animals.

Milk protein genes are transcribed in the mammary gland of lactating animals, and the encoded proteins are secreted in large quantities into milk. The whey acidic protein (WAP) is an abundant milk protein in mice (1, 2) but has not been found in swine or other livestock. Expression of the WAP gene is confined to the mammary gland (2, 3) and is under the control of steroid and peptide hormones as well as other developmental signals during pregnancy (4–6).

By targeting synthesis of foreign proteins to the mammary gland of transgenic animals, it should be possible to produce valuable proteins on a large scale in milk (7, 8). The combined properties of high activity and tissue-specificity make the murine WAP gene promoter a good candidate for targeting gene expression to the mammary gland. Towards this end we previously have expressed a hybrid gene containing regulatory elements from the mouse WAP gene and coding sequences from human tissue plasminogen activator in the mammary gland of transgenic mice (5, 6) and analyzed the protein in milk (5, 9). By characterizing the WAP gene, it may be possible to use its control elements to target expression of hybrid genes in farm animals. However, it is not known whether mammary regulatory elements are gene specific and whether they are functional across species boundaries. In addition, it is not known if the presence of a novel protein may adversely affect the physiology of the mammary gland. To address these questions we introduced the unmodified mouse WAP gene (10) into swine, which themselves do not contain a WAP gene, and analyzed expression of RNA and protein. With this approach, potential problems in interpreting expression data from hybrid genes would not be a factor. Also, potential deleterious physiological effects of a foreign protein might be minimized because the target gene encodes a milk protein that would be confined primarily to the mammary gland.

Swine were chosen for these studies because they offer both economy in animal resources and time when compared to ruminantia as a transgenic animal model and because the questions being addressed did not require harvesting large quantities of milk that would be more easily obtained from dairy animals such as cows, goats, or sheep. The two primary constraints in any large animal transgenic project are the number of fertilized ova obtainable and the number of embryo recipients available. On average it is possible to recover 2–3 times more injectable ova per donor gilt than can be collected from a cow, doe, or ewe. The efficiency of producing expressing transgenic pigs or sheep per injected ovum is about 0.3% (calculated from refs. 11 and 12). Though a live-born-expressing transgenic calf has not been reported, a larger number of ova will probably be required to produce an expressing transgenic cow (13). Furthermore, because swine are polytocous, a recipient sow can carry 5 times as many fetuses as a cow, doe, or ewe. Additionally, the generation interval of swine is ≈11 months, whereas that of goats is between 11 and 21 months and that of cattle at least 24 months. Considering all of these factors, the use of swine rather than cows, goats, or sheep requires one-sixth the number of animals, with results obtainable in less than half the time.

MATERIALS AND METHODS

Production of Transgenic Pigs. Ovulation control and egg recovery were performed as described (14). Briefly, the time of ovulation of sexually mature gilts was controlled by feeding 15 mg of Altrenogest (R-2267, 17-allyl-hydroxyestra-4,9,11-trien-3-one, Roussel-Uclaf) daily for 5–9 days, beginning on day 12 and ending on day 15 of the estrous cycle. Twenty-four hours after the last feeding of Altrenogest, each gilt was given 1000 to 2000 international units of pregnant mare's serum gonadotropin (PMSG) by subcutaneous injection, and 79 hr later each gilt was given an intramuscular injection of 500 international units of human chorionic gonadotropin (hCG). Estrus behavior was monitored, and embryo donor gilts were either bred with a fertile boar or were artificially inseminated with fresh semen twice during estrus.

Approximately 58–61 hr after the hCG injection (18–21 hours after the expected time of ovulation), the reproductive tracts of donor gilts were exposed by midventral laparotomy during general anesthesia. Ova were recovered by flushing 20 ml of Dulbecco's phosphate-buffered saline (15) from the uterotubal junction through the cannulated infundibular end

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: WAP, whey acidic protein.

‡Present address: Hoechst AG, Frankfurt, Federal Republic of Germany.
§To whom reprint requests should be addressed.

of each oviduct. Recovered ova were immediately transferred into BMOC buffer (16) prior to microinjection and maintained at 38°C.

Pig ova are optically opaque and, as a consequence, their nuclear structures are not visible. However, centrifuging ova at ≈15,000 × *g* for 3–8 min displaces the opaque material in the cytoplasm, thereby allowing the nuclear structures to be visualized (14). Pig ova were centrifuged, and a pronucleus of one-celled ova or both nuclei of two-celled ova were injected with a TE solution (1 mM Tris·HCl/0.1 mM EDTA, pH 7.2) containing ≈2 ng of a 7.2-kilobase (kb) *Eco*RI fragment per μl that contained the mouse WAP gene (10). The fragment contained the entire transcribed region with its four exons, three introns, and 2.6-kb 5′ and 1.6-kb 3′ flanking sequences. Microinjections were performed with the aide of differential interference contrast optics at 200-fold magnification, essentially as described for mouse ova (17).

Between 20 and 30 injected ova were deposited into the ampullar region of one oviduct of each recipient gilt whose reproduction cycle had been synchronized with Altrenogest (but not superovulated—i.e., not given PMSG) or whose estrous cycle naturally coincided with the desired stage. Some recipients also received 2–4 uninjected control ova to increase the likelihood of maintaining pregnancy in the event that a majority of the microinjected eggs failed to develop. Time between microinjection and embryo transfer was about 30 min.

To identify transgenic piglets, DNA from tail biopsies was prepared and analyzed for the mouse WAP gene by Southern blotting. Offspring in the F_1 generation were analyzed by the polymerase chain reaction by using primers specific to the WAP gene.

Analysis of Mouse WAP. Milk whey proteins were separated under denaturing conditions in sodium dodecyl sulfate (SDS)/16% polyacrylamide gels and either stained with Coomassie Blue or transferred to nitrocellulose filters. After transfer the membrane was incubated overnight in TBS (20 mM Tris·HCl, pH 7.5/500 mM NaCl) containing 3% gelatin and then was washed in TTBS (TBS containing 0.05% Tween 20). The membrane was then probed for 90 min with a 1:200 dilution of rabbit anti-WAP serum, followed by washing and incubation with alkaline phosphatase-conjugated goat anti-rabbit IgG in TBS containing 1% bovine serum albumin for 1 hr. The antibody–antigen complexes were stained with nitrobluetetrazolium and 5-bromo-4-chloro-3-indolyl phosphate in 100 mM Tris·HCl, pH 9.5/100 mM NaCl/5 mM $MgCl_2$.

Isolation of RNA and Northern Blot Analysis. During necropsy, tissues were immediately placed in liquid nitrogen and stored at −80°C, and total RNA was isolated (18). RNA samples containing 1 μl of ethidium bromide solution (1 mg/ml) were electrophoresed in 1.5% agarose/formaldehyde gels. The gels were blotted onto GeneScreenPlus nylon membranes, which were then probed with a randomly primed labeled 450-base-pair (bp) cDNA fragment that spanned the mouse WAP coding region.

RESULTS

The Mouse WAP Gene in Transgenic Swine. Eight-hundred and fifty ova were recovered and microinjected, of which two-thirds were at the one-cell stage of development. The injected DNA contained 7.2 kb of the mouse WAP gene (see *Materials and Methods*, ref. 10). The microinjected ova along with 34 control ova were transferred into 29 recipient gilts. Twenty-two of the recipients carried their pregnancies to term, resulting in the birth of 189 pigs. DNA analysis of tail biopsies revealed that 5 (2 males and 3 females) of the piglets had incorporated the mouse WAP gene into their genomes. Approximately 1% of the injected ova resulted in transgenic founders. From other transgenic pig projects using different gene constructs, the efficiency of producing founder pigs was similar (11). In this study one pig was stillborn and one died shortly after birth. Such deaths are not uncommon in the pig industry, where neonate mortality is in the range of 15–20%. Lines from the three surviving pigs were established, and offspring were analyzed. Male founder 1301 was bred to three nontransgenic females; 4 of 32 offspring were transgenic, suggesting that he was mosaic for the WAP gene. Transgenic mouse breeding studies have estimated that about 30% of transgenic founders are germ-line mosaics (19). Based on Southern blot analyses, this line contains ≈10 intact copies of the WAP gene in a head-to-tail arrangement at a single locus. Female founder 2202 carried ≈15 copies of the WAP gene. She was bred at 8 months of age; 4 of 9 offspring were transgenic. She was bred a second time and died of an unknown cause 4 days before anticipated parturition. The two transgenic daughters from her first litter were also bred, and after farrowing, milk and RNA were analyzed. Female founder 1302, carrying ≈10 copies of the WAP gene, was unsuccessfully bred three times. After the third failure, she was superovulated as a means of diagnosing the cause of her reproductive failures and to collect eggs if the cause did not involve ovarian disfunction. Twenty-eight ova were recovered and transferred to two recipients. From these, 20 piglets were born of which 8 were transgenic. Apparently not all of female founder 1302's eggs had been recovered because she subsequently gave birth to 9 piglets, 5 of which were transgenic.

Secretion of Mouse WAP into Pig Milk. Expression of the WAP transgene in transgenic pigs was evaluated by both protein and RNA analyses. Milk from female founder 2202 and her daughter 5403, from two daughters (5511 and 5701) of male founder 1301, and from female founder 1302, was analyzed for the presence of mouse WAP. Milk proteins were separated in SDS/polyacrylamide gels and either stained with Coomassie blue or blotted onto nitrocellulose membranes and analyzed with anti-mouse WAP antibodies. WAP has a molecular mass of about 14 kDa (Fig. 1*A*, lane 8) and, at a concentration of about 2 mg per ml, constitutes the major

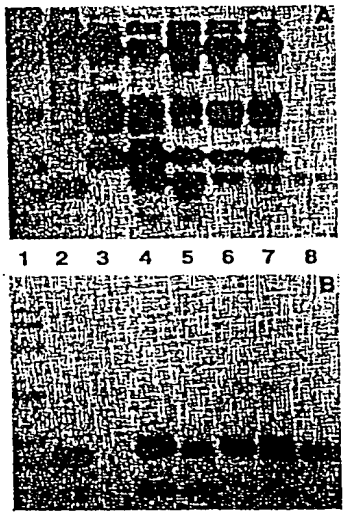


FIG. 1. Secretion of mouse WAP into milk of transgenic pigs. Milk proteins (20 μg) were separated in SDS/polyacrylamide gels and either stained (*A*) or analyzed with rabbit anti-WAP antibodies (*B*). Lanes: 1, molecular mass markers (14, 18, 29, 45, 68, and 96 kDa); 2, total mouse whey proteins; 3–7, milk from nontransgenic pig (lane 3), pig 2202 (lane 4), pig 5403 (lane 5), pig 5701 (lane 6), and pig 5511 (lane 7); 8, 1 μg of purified mouse WAP.

whey protein in mice (Fig. 1*A*, lane 2). A protein comigrating with mouse WAP was present in the milk of transgenic pigs (Fig. 1*A*, lanes 4–7) but not in milk from a nontransgenic control pig (Fig. 1*A*, lane 3). In addition, a 14-kDa protein in milk from transgenic, but not from nontransgenic, pigs reacted strongly with anti-mouse WAP antibodies (Fig. 1*B*). The lower molecular mass material reacting with anti-WAP antibodies probably reflects degradation products of the WAP. Taken together, this shows that the mouse WAP gene was expressed in transgenic pigs, and the encoded protein was secreted into the milk. The level of mouse WAP in the milk of each transgenic pig was determined in ELISA. By setting the level of WAP in mouse milk arbitrarily at 100%, animals 2202 and 5403 (line 2202) and animals 5701 and 5711 (line 1301) were shown to express WAP at about 100%, and female founder 1302, at about 50%. Thus, about 1–2 g of WAP was present per liter of pig milk.

WAP is secreted into mouse milk during the entire lactational period. To determine whether the expression in transgenic pigs paralleled this pattern, we analyzed WAP levels in the milk of founder female 1302 over a 4-week lactational period (Fig. 2). Whey samples were separated in SDS/polyacrylamide gels and either stained (Fig. 2*A*) or analyzed with anti-WAP antibodies (Fig. 2*B*). Constant levels of WAP were found over a 26-day period. This suggests that, at least over this period of time, the WAP transgene was coordinately regulated with other pig milk protein genes.

Expression of Mouse WAP RNA in Pigs. To correlate the level of WAP in milk with the corresponding RNA in mammary tissue, founder female 2202 was biopsied 11 days postpartum, and mammary RNA was analyzed with a mouse-specific WAP cDNA. An RNA of about 600 nucleotides hybridized with the WAP probe (Fig. 3, lanes b and c), confirming mouse WAP gene expression in the mammary glands of transgenic pigs. Furthermore, the RNA levels in pig 2202 and mouse were similar; this agrees with the WAP levels found in the milk. The WAP RNA in pig 2202 appeared to be about 10–20 nucleotides shorter than its counterpart in mice (Fig. 3). Since the protein coding region was intact, the smaller size may be due to differences in polyadenylylation. RNA from a nontransgenic pig did not hybridize with the WAP probe (Fig. 3, lane a), verifying the absence of an endogenous WAP RNA in the pig mammary gland.

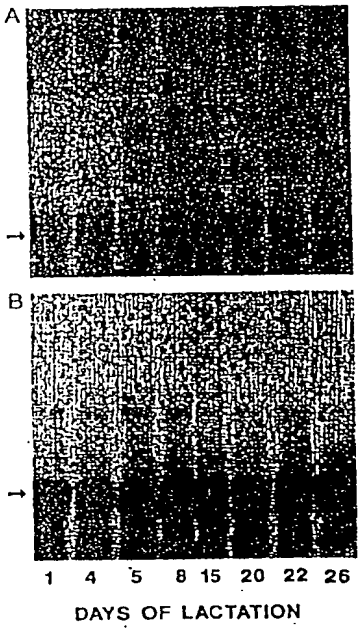


FIG. 2. Expression of mouse WAP during the lactational period of pig 1302. Milk samples were collected at various days after parturition as indicated, and whey fractions were prepared. Upon gel separation, samples were either stained (*A*) or analyzed with anti-WAP antibodies (*B*).

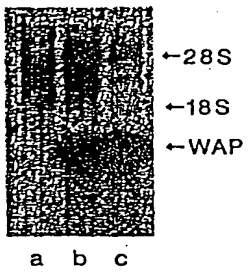


FIG. 3. Expression of mouse WAP RNA in transgenic pigs. Mammary RNAs (5 μg) from a lactating nontransgenic pig (lane a), founder pig 2202 (lane b), and a mouse (lane c) were separated in a formaldehyde gel, transferred to a nylon membrane, and analyzed with a cloned cDNA probe specific for mouse WAP RNA.

In lactating mice the WAP gene is expressed almost exclusively in the mammary gland with levels in nonmammary tissues at least 4 orders of magnitude lower (5). To test whether the 7.2-kb WAP transgene contained elements for stringent tissue specificity observed in mice, we analyzed tissues from lactating pigs from lines 2202 and 1301 for the presence of WAP RNA (Fig. 4). To demonstrate potential WAP expression in nonmammary tissues, we exposed the RNA blot for 24 hr (Fig. 4 *a* and *c*). The specificity of WAP hybridization and the quantity of WAP RNA in the mammary gland were assessed in a 30-min exposure (Fig. 4*b*). In animal 5701 (line 1301), WAP RNA was only found in the mammary gland (Fig. 4*c*) at a level similar to that seen in a 10-day lactating mouse. The sensitivity of the assay would have

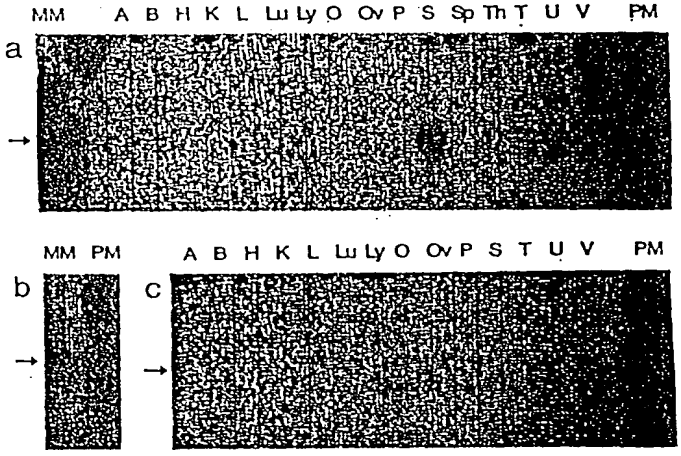


FIG. 4. Tissue distribution of WAP RNA in transgenic pigs. Pigs 5403 (*a*) and 5701 (*c*) were sacrificed, and RNA was prepared from several tissues. Upon separation in formaldehyde gels and transfer to nylon membranes, the RNA was analyzed with a probe specific for mouse WAP RNA. Lanes: MM, mouse mammary gland; PM, pig mammary gland; A, adrenals; B, brain; H, heart; K, kidney; L, liver; Lu, lung; Ly, lymph node; O, ovaries; Ov, oviduct; P, pituitary; S, salivary gland; Sp, spleen; Th, thymus; T, tongue; U, uterus; V, vulva. In *a* and *c*, 20 μg of total RNA was loaded in lanes with the exception of mouse mammary gland (lane MM), where 4 μg was loaded. (*b*) One-hour exposure of the MM and PM lanes of *a*. Arrows indicate the position of WAP RNA.

permitted detection of WAP RNA levels 1000-fold lower than that observed. The level of WAP RNA in animal 5403 (line 2202) was about 80% of that seen in mouse (Fig. 4*a*). The lower molecular mass band in the vulva RNA from animal 5701 was not reproducible and probably reflects a gel or blotting artifact. In animal 5403 WAP expression was detected in salivary gland, although at a level of only 1% of that seen in mammary tissue (Fig. 4*a*). Low-level expression in the salivary gland also has been described for other transgenes containing regulatory elements from milk protein genes (5, 20). Although the salivary gland and mammary gland have similar developmental patterns in that they require interaction between epithelial and mesenchymal tissue for proper duct formation to occur (21, 22), they are not considered closely related. In contrast, sebaceous glands have a common developmental origin to that of the mammary gland. However, no WAP transcripts were found in tissue taken from the vulva (Fig. 4), which is rich in sebaceous glands.

DISCUSSION

Three lines of transgenic swine containing the mouse WAP gene have been generated and analyzed. Although swine does not contain an endogenous WAP gene, its transcription machinery recognized the mouse WAP transgene in a tissue-specific manner, and mouse WAP was secreted into milk from founder swine as well as their offspring at levels similar to those seen in mouse milk. Thus, the molecular basis for mammary-specific gene expression is conserved between swine and mouse, and it can be suggested that the mouse WAP gene shares mammary regulatory elements with pig milk protein genes.

Expression levels of the mouse WAP genes in three lines of transgenic pigs described here and in three additional lines (unpublished data), which carry between 10 and 20 copies of the transgene, were consistently high and at a level comparable to the expression level of the endogenous gene in mice. Activity of the WAP gene in pigs appears to be relatively independent of the site of integration into host chromosomes and also independent of the gene copy number. In contrast, expression of the same 7.2-kb mouse WAP gene in transgenic mice was highly dependent on the integration site of the transgene (36). It remains to be determined whether the consistently high-level expression in transgenic pigs reflects special properties of the WAP gene, such as the presence of dominant transcription elements, or whether the pig genome provides a unique permissive environment for transgene expression. A host of other transgenic swine projects (23) argues against the latter explanation. Data from the sheep β-lactoglobin gene (24), the rat WAP (25) and β-casein (26) genes, and several hybrid genes containing mammary regulatory elements (27–30) have shown that expression was influenced by the site of integration in transgenic mice. At a minimum the present study suggests that WAP gene regulation is different in mice and swine.

This study shows that it is feasible to synthesize and secrete a heterologous milk protein in the milk of farm animals at relatively high concentrations—i.e., more than 1 g/liter. Clark and colleagues had shown that hybrid genes containing regulatory elements from the sheep β-lactoglobulin gene are expressed in the mammary glands of transgenic sheep (31). However, the concentrations of the encoded proteins factor IX and α_1-antitrypsin were only 25 μg/liter and 5 mg/liter, respectively (31). With another transgene, this group produced human α_1-antitrypsin in mouse milk at levels of more than 1 g/liter (20). Therefore, the ability of a transgene to be expressed in the mammary gland at high levels does not appear to be related to the nature of the encoded protein (milk protein versus foreign protein) but rather to the presence of appropriate transcription elements. We are currently testing the ability of the mouse WAP gene promoter to control expression of non-WAP structural gene sequences in pigs.

The concentration of the transgene product produced in this study should be encouraging to those who envision using the mammary gland as a bioreactor for the production of foreign proteins as an economically viable alternative to existing tissue and microbial culture systems (7, 8). Swine produce about 10 kg of milk per day (32), and, based on the expression levels discussed here, it should be possible to produce the protein of interest at a rate of about 1 kg per lactational period of 7 weeks. Since the WAP gene promoter is active in pigs during their entire lactational period, this appears to be an achievable goal, and one sow could satisfy current world's demand of blood clotting facor IX. Alternatively, to the dairy industry, the modification of the composition of milk proteins themselves may be desirable so that overexpressing heterologous or endogenous milk proteins would result in novel milk products (33).

As with other expression systems, high activity of the transgene could have adverse effects on the physiology of the mammary gland. Pigs from two lines (1301 and 2202) were unable to sustain lactation. In contrast, lactation persisted normally in female founder 1302. This animal secreted less WAP into milk than those that abrogated lactation. Agalactia has not been observed in transgenic mice that secrete into their milk heterologous milk proteins (24, 34) or pharmacologically active proteins (20, 35) at levels similar to or exceeding those described here with swine. Experiments are in progress to determine whether the premature termination of lactation exhibited by some of the pigs is associated with mammary gene expression.

Note Added in Proof. We have generated transgenic mice with the 7.2-kb WAP transgene described in this paper and observed that some of the animals cannot maintain lactation (T. Burdon, R.J.W., and L.H., unpublished data).

We thank Floyd Schanbacher for purified mouse WAP, Leah Schulman and Mark Spencer for technical assistance, Jim Piatt for animal care, and William Jakoby for continued support.

1. Piletz, J. E., Heinlen, M. & Ganshow, R. E. (1981) *J. Biol. Chem.* 256, 11509–11516.
2. Hennighausen, L. G. & Sippel, A. E. (1982) *Eur. J. Biochem.* 125, 131–141.
3. Hennighausen, L. G. & Sippel, A. E. (1982) *Nucleic Acids Res.* 10, 2677–2684.
4. Hobbs, A. A., Richards, D. A., Kessler, D. J. & Rosen, J. M. (1982) *J. Biol. Chem.* 257, 3598–3605.
5. Pittius, C. W., Hennighausen, L., Lee, E., Westphal, H., Nichols, E., Vitale, J. & Gordon, K. (1988) *Proc. Natl. Acad. Sci. USA* 85, 5874–5878.
6. Pittius, C. W., Sankaran, S., Topper, Y. & Hennighausen, L. (1988) *Mol. Endocrinol.* 2, 1027–1032.
7. Clark, A. J., Simons, P., Wilmut, I. & Lathe, R. (1987) *Tibtech* 5, 20–24.
8. Hennighausen, L., Ruiz, L. & Wall, R. J. (1990) *Curr. Opinion Biotech.* 1, 74–78.
9. Gordon, K., Lee, E., Vitale, J. A., Smith, A. E., Westphal, H. & Hennighausen, L. (1987) *Bio/Technology* 5, 1183–1187.
10. Campbell, S. M., Rosen, J. M., Hennighausen, L., Strech-Jurk, U. & Sippel, A. E. (1984) *Nucleic Acids Res.* 12, 8685–8697.
11. Pursel, V. G., Hammer, R. E., Bolt, D. J., Palmiter, R. D. & Brinster, R. L. (1990) *J. Reprod. Fertil. Suppl.* 41, 77–87.
12. Rexroad, C. E., Hammer, R. E., Behringer, R. R., Palmiter, R. D. & Brinster, R. L. (1990) *J. Reprod. Fertil. Suppl.* 41, 119–124.
13. Massey, J. M. (1990) *J. Reprod. Fertil. Suppl.* 41, 199–208.
14. Wall, R. J., Pursel, V. G., Hammer, R. E. & Brinster, R. L. (1985) *Biol. Reprod.* 32, 645–651.
15. Dulbecco, R. & Vogt, M. (1954) *J. Exp. Med.* 99, 167–175.

16. Brinster, R. L. (1972) in *Growth, Nutrition, and Metabolism of Cells in Culture*, eds. Rothblatt, G. & Cristofalo, V. (Academic, New York), Vol. 2, pp. 251–286.
17. Gordon, J. W., Scangos, G. A., Plotkin, D. J., Barbosa, J. A. & Ruddle, F. H. (1980) *Proc. Natl. Acad. Sci. USA* 77, 7380–7384.
18. Chomczynski, P. & Sacchi, N. (1987) *Anal. Biochem.* 162, 156–159.
19. Wilkie, T. M., Brinster, R. L. & Palmiter, R. D. (1986) *Dev. Biol.* 118, 9–18.
20. Archibald, A. L., McClenaghan, M., Hornsey, V., Simons, J. P. & Clark, A. J. (1990) *Proc. Natl. Acad. Sci. USA* 87, 5178–5182.
21. Mepham, T. B. (1987) in *Physiology of Lactation* (Open Univ. Press, Philadelphia), pp. 1–14.
22. Hopper, A. F. & Hart, N. H. (1985) in *Foundations of Animal Development* (Oxford Univ. Press, New York), 2nd Ed., pp. 302–337.
23. Pursel, V. G., Pinkert, C. A., Miller, K. F., Bolt, D. J., Campbell, R. G., Palmiter, R. D., Brinster, R. L. & Hammer, R. E. (1989) *Science* 244, 1281–1288.
24. Simons, J. P., McClenaghan, M. & Clark, A. J. (1987) *Nature (London)* 328, 530–532.
25. Bayna, E. M. & Rosen, J. M. (1990) *Nucleic Acids Res.* 18, 2977–2985.
26. Lee, K.-F., DeMayo, F. J., Atiee, S. H. & Rosen, J. M. (1988) *Nucleic Acids Res.* 16, 1027–1041.
27. Shu-Hua, Y., Deen, K. C., Lee, E., Hennighausen, L., Sweet, R. W., Rosenberg, M. & Westphal, H. (1989) *Mol. Biol. Med.* 6, 255–261.
28. Andres, A.-C., Schönenberger, C.-A., Groner, B., Hennighausen, L., LeMeur, M. & Gerlinger, P. (1987) *Proc. Natl. Acad. Sci. USA* 84, 1299–1303.
29. Lee, K.-F., Atiee, S. H. & Rosen, J. M. (1989) *Mol. Cell. Biol.* 9, 560–565.
30. Bühler, T. A., Bruyere, T., Went, D. F., Stranzinger, G. & Bürki, K. (1990) *Bio/Technology* 8, 140–146.
31. Clark, A. J., Bessos, H., Bishop, J. O., Brown, P., Harris, S., Lathe, R., McClenaghan, M., Prowse, C., Simons, J. P., Whitelaw, C. B. A. & Wilmut, I. (1989) *Bio/Technology* 7, 487–492.
32. Harkins, M., Boyd, R. D. & Bauman, D. (1989) *J. Anim. Sci.* 67, 1997–2008.
33. Hennighausen, L. (1990) *Protein Expres. Purif.* 1, 3–8.
34. Vilotte, J.-L., Soulier, S., Sinnakre, M.-G., Massoud, M. & Mercier, J.-C. (1989) *Eur. J. Biochem.* 186, 43–48.
35. Meade, H., Gates, L., Lacy, E. & Lonberg, N. (1990) *Bio/Technology* 8, 443–446.
36. Burdon, T., Wall, R. J., Sankaran, L. & Hennighausen, L. (1991) *J. Biol. Chem.*, in press.

the business of xenotransplantation

past and present

home
four patients
risks
animal welfare
the business
regulators

In early 1996, an analyst at Salomon Brothers investment firm, detailed like never before the "unrecognised potential of xenotransplantation": a $6 billion market in transgenic organs by 2010. The report was read by big and small investors alike--biotech venture capitalists (who pumped money into xenotransplant research), as well as newspaper personal investment columnists who featured companies like Imutran, Nextran, Alexion, and BioTransplant as "hot picks." In these heady times, some companies were even suggesting that we might each have our own Astrids, "self pigs" custom-made from our own DNA, "immunological twins" available for any spare parts we might need in the course of our lives.

Some five years later, the big profits have not yet been realized. In late 2000, a number of the original players in the xeno business reorganized their efforts for the next phase of research and development: Most notably,Novartis, the Swiss pharmaceutical giant, merged Imutran with BioTransplant to form a new company, Immerge BioTherapeutics; the new company is allied with another Novartis-funded company, Infigen, for use of Infigen's patented cloning technology. While this move was said to reflect a new commitment to xenotransplanation by Novartis, another company's "refocusing" seemed to start with a vote of no confidence for pig-to-human transplants: In August of 2000, PPL Therapeutics, a Scottish company that set out to commercialize the "Dolly" cloning technology, lost "considerable funding" for its xenotransplantation program from Geron, a California-based company that had been PPL's largest xeno backer. PPL and Geron both denied that the move should be construed as any kind of judgement on the viability of pig-to-human transplants, but PPL has had difficulty finding a new partner for its xeno program.

Here's a look at the major players in the xeno business, past and present:

imutran

This small biotech start-up in Cambridge, England took the early lead in the race toward the organ farm: In December of 1992, at a farm in

Cambridgeshire, they created "Astrid," the world's first transgenic pig, who carried human genes within her organs to help prevent rejection by the organ recipient's immune system (one of the thorniest problems facing xenotransplantation). A year and half later, Imutran announced it had produced several generations of Astrids who might be eligible for human trials by 1996. Started by a handful of scientists, the company received early funding from Sandoz pharmaceutical company whose profits were heavily derived from an immunosuppressive drug key to successful transplants. Sandoz eventually purchased Imutran outright. In 1996, Sandoz merged with Ciba-Geigy to form Novartis.

In January, 2001, Imutran completed the relocation of its xenotransplantation research to Charlestown, MA, combining with BioTransplant (another Novartis-funded company) to form Immerge BioTherapeutics. The company denies that the move was motivated by animal rights protests in the UK.

nextran

In the early 1990's, a small biotech company in Princeton, NJ--the DNX Corporation--emerged as one of Imutran's chief rivals. Using a farm in Albany, Ohio, Nextran successfully produced transgenic pigs whose hearts survived for impressive lengths of time in baboons; they were also far along in developing pig livers as filter "bridge" organs for people awaiting transplants. In late August, 1994, the Baxter Health Care Corp. of Deerfield IL partnered with DNX to form a new company--NEXTRAN--with Baxter owning 70% of the partnership. At the time of the formation of Nextran, Baxter's biggest revenue-generator had come from its dialysis equipment, so it took a special interest in DNX, which had been developing transgenic kidneys that might one day make dialysis less necessary. In 1995, Nextran became the first to win FDA approval for human clinical trials involving transgenic pig livers.

In FRONTLINE's report, a Nextran pig saved Robert Pennington's life. It was used outside his body as a temporary "bridge" to filter Pennington's blood while he waited for a human liver transplant. Nextran also is involved in trying to solve the problem of hyperacute rejection problems facing pig-to-human transplants.

alexion

Formed in 1992 by a group of Yale University scientists, Alexion was one of the early innovators in finding transgenic solutions to hyperacute rejection in transplant organs. Though initially focused on creating organs (their pigs were grown on farms in West Virginia and Massachusetts, Alexion has had some of its greatest success with implantation of pig nerve cells to repair spinal cord damage. In late

1998, Alexion made headlines worldwide for successfully repairing severed spinal cords in rats and monkeys using pig cells. Alexion's clinical trials continue.

ppl therapeutics

Based in Edinburgh, PPL Therapeutics is licensed to commercialize the cloning technology pioneered by the Roslin Institute which surprised the world in 1997 with its creation of "Dolly," the first cloned mammal. In 1998, the company moved its xenotransplantation program to Blacksburg, West Virginia where scientists affiliated with Virginia Tech University were already involved in the research. In March of 2000, PPL's Blacksburg laboratory announced the creation of the world's first cloned pigs. (Later in the year, Wisconsin-based Infigen would be the first to clone *transgenic* pigs; these pigs were first shown nationally in FRONTLINE's report.)

In August of 2000, PPL Therapeutics' xenotransplantation program lost "considerable funding" from its major backer, Geron corporation of Northern California, who cited a change in "strategic priorities" and a desire to concentrate on stem cell work. PPL executives as well as the director of Edinburgh's Roslin Institute issued press releases denying that the Geron move was a vote of no confidence for xeno: "The institute has had a research programme on pig cloning, one application of which would be the use of pig organs for xenotransplantation. While xeno has raised a number of well-publicised issues, such as possible infection with pig viruses, these were not the basis for the decision to refocus the funding."). PPL continued to look for partners through the Fall of 2000, but negotiations broke down, largely due to questions about the value of PPL's xeno program.

In early 2001, PPL's Blacksburg, VA lab announced that it had secured new funding--not for xeno, but for stem cell research. It's too soon to tell whether this is one company's story, or a cautionary tale for the industry.

biotransplant

Founded in 1990 and taken public with a stock offering in 1996, BioTransplant was one of the early pioneers of xenotransplantation. Like their rivals, BioTransplant focused on overcoming the hyperacute rejection problem, basing their approach on the bone marrow research of Dr. David Sachs. In August, 2000, the company, which is partnered with Massachusetts General Hospital, announced a breakthrough in breeding transgenic pigs that would not transmit pig viruses, or PERV's.

In January of 2001, BioTransplant spun-off its xenotransplantation

program, partnering with Novartis (and the former Imutran) in a new company, Immerge BioTherapeutics, but keeping their offices in the Charlestown Naval Shipyard.

infigen

Infigen was created in 1997 to commercialize the animal cloning techniques developed at American Breeder Service (ABS Global Inc.)--a DeForest Illinois company which is part of W.R. Grace. (ABS describes itself as "the world's leading provider of bovine reproductive services and technologies," a global marketer of dairy and beef cattle semen.) In January of 1999, Infigen and Imutran (Novartis) formed a working alliance that guaranteed Infigen's funding in exchange for use of the company's patented nuclear transfer cloning techniques.

In his FRONTLINE interview,Michael Bishop PhD Infigen's president, explains how genetically modified pigs can be created and cloned.

diacrin

Founded in 1990, Diacrin became a public company in early 1996, after the FDA gave the company approval for the first-ever clinical trials of transplanted pig cells into humans. Later in 1996, Diacrin entered a joint venture with Genzyme to develop two products using pig neural cells. .

On March 16, 2001 Genzyme and Diacrin reported that a preliminary analysis of outcomes of Phase II trials for Parkinson's patients found pig neuro cell transplants did not necessarily work better than a placebo treatment. The results are likely not the end of the research trial, but the news triggered a significant drop in stock prices. Jim Finn, a Parkinson's patient featured in FRONTLINE's report, was part of a Phase I trial. Other Diacrin/Genzyme Phase I patients featured in FRONTLINE's report--Maribeth Cook and Amanda Davis--were stroke patients.

immerge biotherapeutics

Beginning operations in January of 2001, Immerge BioTherapeutics is a new company formed from the UK's Imutran and the xeno division of the Boston-based BioTransplant company. Unlike the companies from which it was formed, Immerge is focused squarely on development of cells, tissues, and organs for xenotransplantation, and not on drug therapies or other transgenics.

EXHIBIT HH

About Entrez
Text Version

Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities

PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
LinkOut
My NCBI (Cubby)

Related Resources
Order Documents
NLM Catalog
NLM Gateway
TOXNET
Consumer Health
Clinical Alerts
ClinicalTrials.gov
PubMed Central

1: Virology. 1987 Mar;157(1):236-40.

Related Articles, Links

Transgenic chickens: insertion of retroviral genes into the chicken germ line.

Salter DW, Smith EJ, Hughes SH, Wright SE, Crittenden LB.

We infected early chicken embryos by injection of wild-type and recombinant avian leukosis viruses into the yolk of unincubated, fertile eggs. The viremic males (designated generation 0 (G-0] were tested for transmission of proviral DNA to their G-1 progeny. Nine of 37 G-0 viremic males were mosiac and proviral DNA was transmitted to their progeny at frequencies varying from 1 to 11%. All of the G-1 progeny examined by restriction enzyme analysis for clonality of proviral junction fragments had one to three simple but different fragments. The proviral DNA was transmitted from G-1 to the G-2 progeny in a Mendelian fashion thus proving that retroviral genes have been inserted into the chicken germ line. One of the viruses is a candidate vector for insertion of foreign genes into the chicken germ line.

PMID: 3029962 [PubMed - indexed for MEDLINE]

EXHIBIT II

JOURNAL OF VIROLOGY, June 1989, p. 2680–2689
0022-538X/89/062680-10$02.00/0

Vol. 63, No. 6

Replication-Defective Vectors of Reticuloendotheliosis Virus Transduce Exogenous Genes into Somatic Stem Cells of the Unincubated Chicken Embryo

ROBERT A. BOSSELMAN,[1]* ROU-YIN HSU,[1] TINA BOGGS,[1] SYLVIA HU,[1] JOAN BRUSZEWSKI,[1] SUSAN OU,[1]† LARRY SOUZA,[1] LEE KOZAR,[1] FRANK MARTIN,[1] MARGERY NICOLSON,[1] WILLIAM RISHELL,[2] JOSEPH A. SCHULTZ,[2] KENNETH M. SEMON,[2] AND R. GREGORY STEWART[2]‡

Amgen Inc., 1900 Oak Terrace Lane, Thousand Oaks, California 91320,[1] and Arbor Acres Farm, Inc., Marlborough Road, Glastonbury, Connecticut 06033[2]

Received 26 September 1988/Accepted 24 February 1989

Replication-defective vectors derived from reticuloendotheliosis virus were used to transduce exogenous genes into early somatic stem cells of the chicken embryo. One of these vectors transduced and expressed the chicken growth hormone coding sequence. The helper cell line, C3, was used to generate stocks of vector containing about 10^4 transducing units per ml. Injection of 5- to 20-μl volumes of vector directly beneath the blastoderm of unincubated chicken embryos led to infection of somatic stem cells. Infected embryos and adults contained unrearranged integrated proviral DNAs. Embryos expressed the transduced chicken growth hormone gene and contained high levels of serum growth hormone. Blood, brain, muscle, testis, and semen contained from individuals injected as embryos contained vector DNA. Replication-defective vectors of the reticuloendotheliosis virus transduced exogenous genes into chicken embryonic stem cells in vivo.

Insertion of genetic information into the chicken provides a new in vivo approach to analyzing gene expression and its effects on avian physiology. A vector derived from Rous sarcoma virus has been used to transfer additional growth hormone genes into chicken somatic cells by infection of 7- and 9-day-old embryos (35). More recently, gene transfer into chicken germ cells (27–29) has been accomplished by infection of day-old embryos with similar replicating Rous sarcoma virus vectors (18, 33). This approach to avian gene transfer has advantages over DNA microinjection since the early chicken zygote is difficult to manipulate and even a freshly laid egg contains thousands of cells (10, 20). However, replicating retroviral vectors have disadvantages. They can result in gene transfer to susceptible cells at various stages of differentiation long after initial infection of the embryo. This can make it difficult to determine the stage of development at which gene insertion takes place or the cell lineage relationships within fully differentiated tissues. Furthermore, replicating vectors also increase the potential for disease states associated with chronic viral infection (16, 24, 38).

Replication-defective retroviral vectors offer an alternative approach (2, 6, 21, 36, 40). Such vectors, derived from reticuloendotheliosis virus type A (REV-A) (31), are produced by the helper cell line C3 which contains a packaging-defective helper provirus (40). When transfected with a defective proviral vector, this helper cell assembles infectious replication-defective vector but little or no competent virus (17). Both replicating REV and the replication-defective REV vector ME111 have been previously used for gene transfer into chicken somatic cells by injection of virus into follicles before ovulation (32). We have used a method of gene transfer based on microinjection of vector into early embryos.

This report describes the transfer of new genetic information, including additional chicken growth hormone (cGH) coding sequences, into somatic stem cells of the chicken embryo. Chickens do not generally contain endogenous REV and express endogenous cGH only in the pituitary during late embryogenesis and after hatching (15, 19, 30). In vivo, these vectors can infect somatic stem cells of day-old chicken embryos, resulting in precociously high levels of circulating cGH and the presence of vector DNA in a variety of adult somatic tissues.

MATERIALS AND METHODS

Cells. The REV-A helper cell line, C3, was generously provided by H. Temin (40). C3 cells were cultured in minimal essential medium (Eagle) containing 7% fetal calf serum–400 μg of G418 per ml. Chicken embryo fibroblasts (CEF) were grown in F-10 medium supplemented with 10% tryptose phosphate broth–5% calf serum. D17 cells were cultured in minimum essential medium (Eagle)–7% fetal calf serum (40). Buffalo rat liver thymidine kinase (TK)-negative ($BRLtk^-$) cells were grown in minimum essential medium (Eagle) plus 7% calf serum (39). QT-6 cells were obtained from C. Moscovici and grown as described previously (23).

Virus infection. $BRLtk^-$ cells were infected in medium containing 100 μg of Polybrene per ml. CEF were infected in normal medium. Cells were usually exposed to virus overnight.

Vectors. The ME111 vector has been previously described (8). The vector SW272/cGH was derived by insertion of cGH cDNA downstream of the 5' long terminal repeat (LTR) of the SW272 vector (39).

Vector assays. TK transducing units (TKTU) released by 5×10^5 C3 helper cells stably transfected with vector SW272/cGH were harvested after 6 h of incubation and were assayed by infection of 10^5 $BRLtk^-$ cells. TK-positive cells were selected for growth in medium (40) containing 1×10^{-4}

* Corresponding author.
† Present address: Department of Biology, California Institute of Technology, Pasadena, CA 91125.
‡ Present address: Poultry Research Center, University of Georgia, Athens, GA 30601.

M hypoxanthine, 3×10^{-6} M thymidine, and 5×10^{-7} M methotrexate.

Plasmids and chicken DNAs. All plasmid DNAs were propagated by using derivatives of pBR322 and the HB101 strain of *Escherichia coli*. The plasmid pSW272 contains a derivative of the spleen necrosis provirus (SNV), lacks most sequences encoding the structural genes of the virus, and contains the herpes simplex virus type 1 (HSV-1) *tk* gene (40). Chicken genomic DNAs were isolated from Arbor Acres males of meat breeding lines.

Nucleic acid isolation. Chicken embryo DNA was prepared by solubilizing tissue in buffer containing 100 mM EDTA, 1% sodium dodecyl sulfate, 100 μg of proteinase K per ml (pH 8). Samples were incubated at 60°C for 15 min, then at 37°C with additional protease (100 μg/ml) for 4 h. The DNA was sheared, adjusted to 200 mM NaCl, and extracted twice with equal volumes of phenol and chloroform-isoamyl alcohol (24:1) and once with 2 volumes of chloroform-isoamyl alcohol. DNA was ethanol precipitated and dissolved in 0.01 M Tris–0.001 M EDTA (pH 8.0). Unsheared DNA was used for Southern blot analysis (34).

Nucleic acid analysis. DNA samples were applied to Gene Screen Plus membranes (New England Nuclear Co.) for dot blot analysis by means of 96-well plexiglass manifolds. DNA on membranes was denatured in 1.5 M NaCl–0.5 M NaOH for 15 min, neutralized in 0.5 M Tris(pH 7.5)–1.5 M NaCl for 1 min, blotted dry, and baked at 80°C for 30 min. Hybridizations were carried out as already described (17). Radiolabled DNA probe was prepared by the method of random priming (13). Southern blot analysis was performed as described previously (34).

cGH analysis. cGH expression was analyzed either by radioimmunoassay (RIA) (35) or by Western immunoblotting (3).

Transfection. The REV-derived helper cell line, C3, was transfected as previously described (14) with the plasmids pSW272/cGH and pHyG (37). Transfected cells were selected for 10 to 14 days in medium containing 200 μg of hygromycin per ml.

Embryo infection. Shell was removed from the area above the blastoderm of unincubated eggs. A Narishigi micromanipulator and a 25-μl Drummond pipette fitted with a glass needle were used to inject 5- to 20-μl volumes of cell culture medium containing vector directly beneath the exposed blastoderm. The titer of vector was about 10^4 TKTU/ml as measured on BRLtk$^-$ cells. The relative titer of this vector on chicken embryo cells in vivo is unknown. Eggs were resealed with a patch of shell membrane which was covered with Devcon Duco cement and allowed to dry. Eggs were incubated at 37.8°C.

RESULTS

Vectors ME111 and SW272/cGH. The sequence relationships among SNV, ME111, the cGH transducing vector SW272/cGH, and the packaging-defective helper proviruses present in C3 helper cells are shown in Fig. 1. ME111 has been described in detail elsewhere (8). The parental vector SW272 is derived from SNV and contains the HSV-1 *tk* gene and promoter in the same transcriptional orientation as the viral promoter (39). The cGH coding sequence was originally derived from a cDNA clone made from chicken pituitary mRNA (35). A DNA fragment *Xba*I to *Nco*I contains the complete coding sequence of the cGH gene but lacks the poly(A) addition signal present at the 3′ end of the cDNA. Using Klenow reagent and blunt-end ligation, the cGH sequences were inserted into the unique *Xba*I site within pSW272 located just downstream of the viral 5′ splice donor and packaging sequence, 555 nucleotides from the 5′ end of the viral RNA transcript (39). The orientation of the cGH coding sequence is the same as that of the viral sequences. Proceeding from the 5′ end of the proviral RNA transcript of SW272/cGH, the first ATG encountered codes for the N-terminal methionine of cGH. SW272/cGH is designed to express cGH mRNA transcripts from the viral promoter.

Transduction and expression of REV vectors in vitro. Careful screening of the C3 helper cells transfected with pSW272/cGH and pHyg yielded clone C3-44 which released 2×10^4 TKTU/ml into growth medium but very low levels of competent virus. Competent REV in these cultures, as estimated by infection of cultured CEF, was about 10 infectious units of REV per ml or less (17). Western blot analysis of cGH released by C3-44 cells revealed a predominantly single band of protein which comigrated with purified recombinant cGH (Fig. 2). The observed molecular size of cGH was about 23,000 daltons. The estimated concentration of cGH in a 72-h harvest of medium of clone C3-44 was at least 500 ng/ml (data not shown). CEF infected with vector released >40 ng of cGH per ml of growth medium as determined by RIA 3 days after infection (data not shown). Western blot analyses of cGH released by cell lines infected with the SW272/cGH vector are shown in Fig. 2, lanes 13 through 18. Cell lines B56 and B20 derive from the canine cell line D17. Cell lines QT82, QT54, QT15, and QT8 derive from the quail cell line QT-6. All of these cells release cGH having the same apparent molecular size as purified recombinant cGH (23 kilodaltons) (35). Approximate levels of cGH expression varied from 2 to 10 ng/ml.

Analysis of DNA from chicken embryos after vector infection. Tissue culture fluid (20-μl volumes) containing the vector SW272/cGH was injected beneath the blastoderms of unincubated chicken embryos. Total embryonic DNA was isolated from vector-injected and uninjected control embryos after 7 days of development and was analyzed by qualitative dot blot hybridization with either a radiolabeled cGH probe (Fig. 3A) or a REV vector probe (Fig. 3B). The cGH probe was used to demonstrate that sufficient DNA was present on the filter for detection of vector sequences present at low copy number. Of 25 injected embryos, 13 (52%) hybridized to a radiolabeled probe of vector DNA, whereas control DNA from uninjected embryos did not.

To confirm the presence and correct genome organization of vector sequences in infected 7-day embryos, high-molecular-size DNAs from 10 vector-containing embryos were digested with *Bam*HI endonuclease and subjected to Southern blot analysis (34) (Fig. 4). The embryo DNAs examined included those from Fig. 3B, rows 1a, 2a, 6a, 7a, and 8a. Internal *Bam*HI fragments predicted from the cGH vector sequence are diagrammed in Fig. 1. Digestion of integrated proviral vector sequences of SW272/cGH should yield DNA fragments internal to the provirus of 0.86, 2.3, and 1.6 kilobase pairs (kb). A 5′ junction fragment containing the 5′ LTR of the vector linked to host cellular sequences adjacent to the integration site might also be detected. No 3′ junction fragment containing host DNA sequences would be detected, because a *Bam*HI restriction endonuclease site is located at the 3′ end of the proviral LTR. As shown in Fig. 4A, lanes 3 to 7 and 12 to 16, DNAs from these vector-infected embryos show the expected *Bam*HI DNA fragments of 0.86, 2.3, and 1.6 kb when analyzed with a probe derived from the complete SW272 plasmid DNA, which does not contain cGH sequences. The absence of detectable *Bam*HI fragments containing the junction of cellular DNA and integrated vector DNA indicates multiple sites of vector

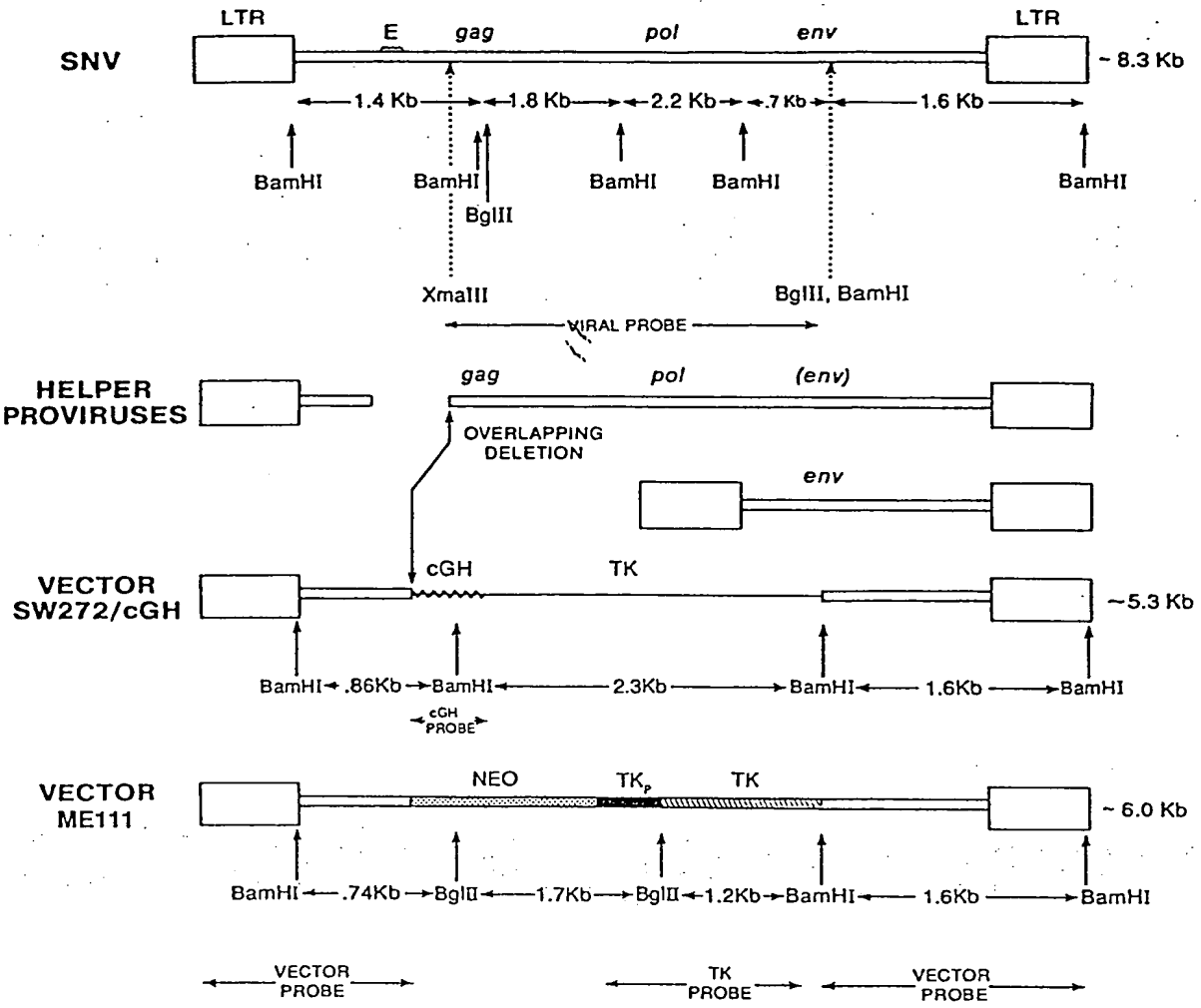


FIG. 1. Sequence relationships among the parental SNV provirus, the modified packaging-defective helper proviruses, and the vectors ME111 and SW272/cGH are shown. Relevant features of these proviruses include the LTRs, the structural genes of the virus (*gag*, *pol*, *env*), the approximate position of the packaging sequence (E), the cGH sequences, the HSV-1 *tk* gene promoter (TKp), the *tk* coding sequence (TK), and the neomycin phosphotransferase coding sequence (NEO). The (*env*) sequence in the larger of the two helper proviruses is presumably not expressed because of removal of the 5' splice donor. Overlapping deletions indicated between helper and vector sequences should reduce recombination between these genomes. A description of the REV helper proviruses and the original TK transducing vector pSW272 and ME111 have been given (8, 40). The 5' LTRs of both helper proviruses derive from SNV. Their coding sequences derive from REV-A. The *env* helper provirus lacks viral splice donor and acceptor sequences. The first ATG is that of the *env* gene. The cGH vector derives from SNV. REV-A and SNV share high sequence homology. Relative sizes (in kilobases) of *Bam*HI restriction endonuclease fragments are indicated. Also given are the locations of viral, vector, TK, and cGH DNA probes.

provirus integration during infection of early embryonic cells. No 0.57-kb *Bam*HI fragment predicted from the structure of unintegrated circular forms of either the vector DNA or helper virus DNA was observed. No 1.4-kb fragment diagnostic of the 5' end of integrated replication-competent proviral SNV DNA was observed (39) (see Fig. 1). *Bam*HI-digested DNA from uninjected whole embryos or from blood of uninjected chickens did not hybridize to the vector probe (Fig. 4A, lanes 2, 8, 11, and 17, respectively).

After removal of the SW272 probe (Fig. 4B), the same filters were hybridized with a viral probe specific for the structural genes of REV to detect the presence of replication-competent virus (Fig. 4C). The parental SNV and REV-A proviruses used to derive the helper cell and vectors described here contain internal *Bam*HI fragments of 1.4, 1.8, 2.2, 0.7, and 1.6 kb (see Fig. 1). Only the 1.6-kb fragment would not be detected by the virus-specific probe (Fig. 1) used in this analysis. No virus-specific *Bam*HI fragments were observed, indicating that endogenous and exogenous REV sequences were not detectable (Fig. 4C). Although this result does not rule out the presence of competent helper virus, it shows that efficient gene transfer takes place via the replication-defective SW272/cGH vector. The dot blot on the right of panel C contains various quantities of plasmid pSW253 which carries the entire REV provirus (5).

The filters shown in Fig. 4C were washed to remove probe (Fig. 4D) and were reanalyzed with a cGH-specific probe (Fig. 4E). The fragments of 0.86 and 2.3 kb in lanes 3 to 7 and 12 to 16 are the predicted cGH-containing vector sequences described in Fig. 1. The two bands (asterisks) of approxi-

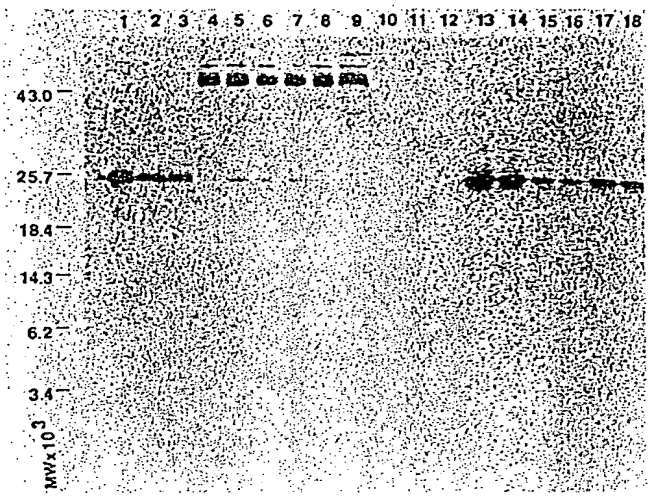


FIG. 2. Western blot analysis of cGH. Lanes: 1 to 3, 250, 125, and 30 ng per lane, respectively, of purified recombinant cGH; 4, 10 μl of conditioned QT-6 medium; 5 to 9, immunoprecipitated cGH from various volumes of serum from 15-day vector-injected embryos, birds 18 (80 μl), 2 (100 μl), 6 (60 μl), 30 (100 μl), and 27 (200 μl), respectively (see Table 1); 10 to 12, sample buffer alone; 13 to 18, 10 μl of conditioned medium from clones of SW272/cGH-infected D17 and QT-6 cells (clones B56, B20, QT82, QT54, QT15, and QT8, respectively). Molecular weights (MW × 10^3) are shown at the left.

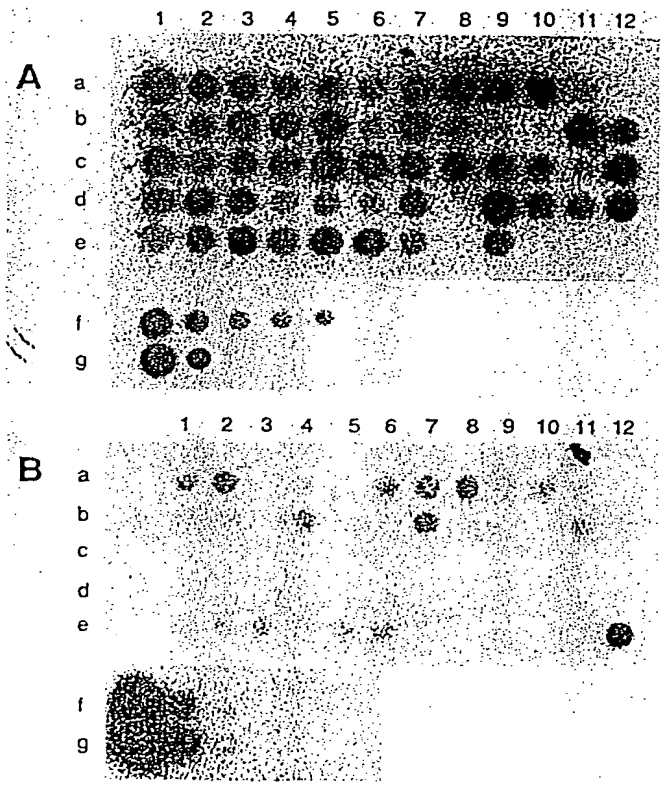


FIG. 3. Dot blot hybridization analysis of chicken embryo DNA. (A) Chicken embryo DNA hybridized with a cGH probe. Rows: a, b, and c1, total DNA from 7-day embryos injected with 20 μl of vector C3-44; c2 to c12 and d1 to d7 and d9, DNA from uninjected control embryos; d10 to d12, e1 to e3, e4 to e6 and e7 to e9, DNA from brain, liver, and muscle of four 14-day embryos injected with vector SW272/cGH; f1 to f4 chicken blood DNA (~15 μg) mixed with 1/10 dilutions of vector DNA starting with 1 ng in spot f1; f5, chicken blood DNA only; g1 to g4, yeast tRNA (5 μg) mixed with the same amounts of vector DNA present in rows f1 to f4; g5, yeast tRNA only. (B) Same as in panel A, except filters were hybridized with a vector-specific probe (see Fig. 1). Row e12 is the same as row a1. Approximately 15 to 30 μg of total embryo DNA was applied to each spot, using a 96-well blotting apparatus.

mately 6.4 kb and approximately 2.7 kb, which are common to all lanes, represent *Bam*HI fragments derived from the endogenous cGH gene. As expected, embryo DNAs in lanes 3 to 7 and 12 to 16 contain all four fragments derived from both the vector and endogenous gene. The 1.6-kb *Bam*HI fragment present in lanes 3 to 7 and 12 to 16 of Fig. 4A is missing in Fig. 4E, because this fragment does not contain cGH sequences.

Dot blot hybridization of DNA from brain, liver, and muscle of four 14-day embryos infected before incubation showed that two of the four embryos contained vector-specific sequences in all three tissues. One embryo contained vector sequences in liver and muscle only, and one embryo was negative (Fig. 3).

Analysis of serum cGH. Circulating levels of cGH were determined by RIA of serum from thirty 15-day-old embryos infected with vector before incubation (Table 1). Concentrations of cGH in serum from 16 of 30 injected embryos (55%) were at least 10 times the level in uninjected control embryos, and they ranged from 18 to 254 ng/ml. All 35 control embryos contained less than 2 ng of detectable serum cGH per ml. Western blot analysis of cGH immunoprecipitated from serum of a number of these embryos is shown in Fig. 2, lanes 5 to 9. The amount of cGH present in serum from infected embryos is similar to the amount of cGH produced in vitro by infected culture cells.

Vector sequences in adult chickens. Southern blot analysis of DNA isolated from blood, brain, muscle, and testis of an adult chicken (no. 87725) which had been injected as an embryo with the ME111 vector is shown in Fig. 5. DNAs were digested with *Bam*HI and *Bgl*II before analysis. The four different probes used hybridized with the REV sequence present in the vector, HSV-1 *tk* sequences of the vector, REV structural gene sequences (absent from the vector), or endogenous cGH genes. All analyzed DNAs from bird 87725 contained the predicted DNA fragments of 0.74 and 1.6 kb recognized by the REV vector probe and fragments of 1.2 and 1.7 kb recognized by the *tk* probe (Fig. 5A and C, respectively). DNA from blood and brain contained additional hybridizing fragments which probably include junctions between vector and cellular DNA at sites of integration (Fig. 5A, lanes 2 and 3). No REV-specific bands were observed in any of these tissue DNAs (Fig. 5B). Hybridization with cGH probe revealed endogenous fragments of ~2.7 and ~6.4 kb (Fig. 5D). D17 cells cocultivated with blood taken from bird 87725 at 4 weeks of age were reverse transcriptase-negative after 4 weeks of culture and did not produce detectable *tk* gene-transducing activity. Of 14 similarly derived birds, 2 were virus positive as determined by the same assay (17). Although the presence of low levels of replicating REV in birds like no. 87725 cannot be ruled out, these results are consistent with infection of embryonic stem cells with nonreplicating REV vectors.

Southern blot analysis of DNA from semen and blood of SW272/cGH-positive and control birds is shown in Fig. 6. Filters containing *Bam*HI-digested DNAs were hybridized

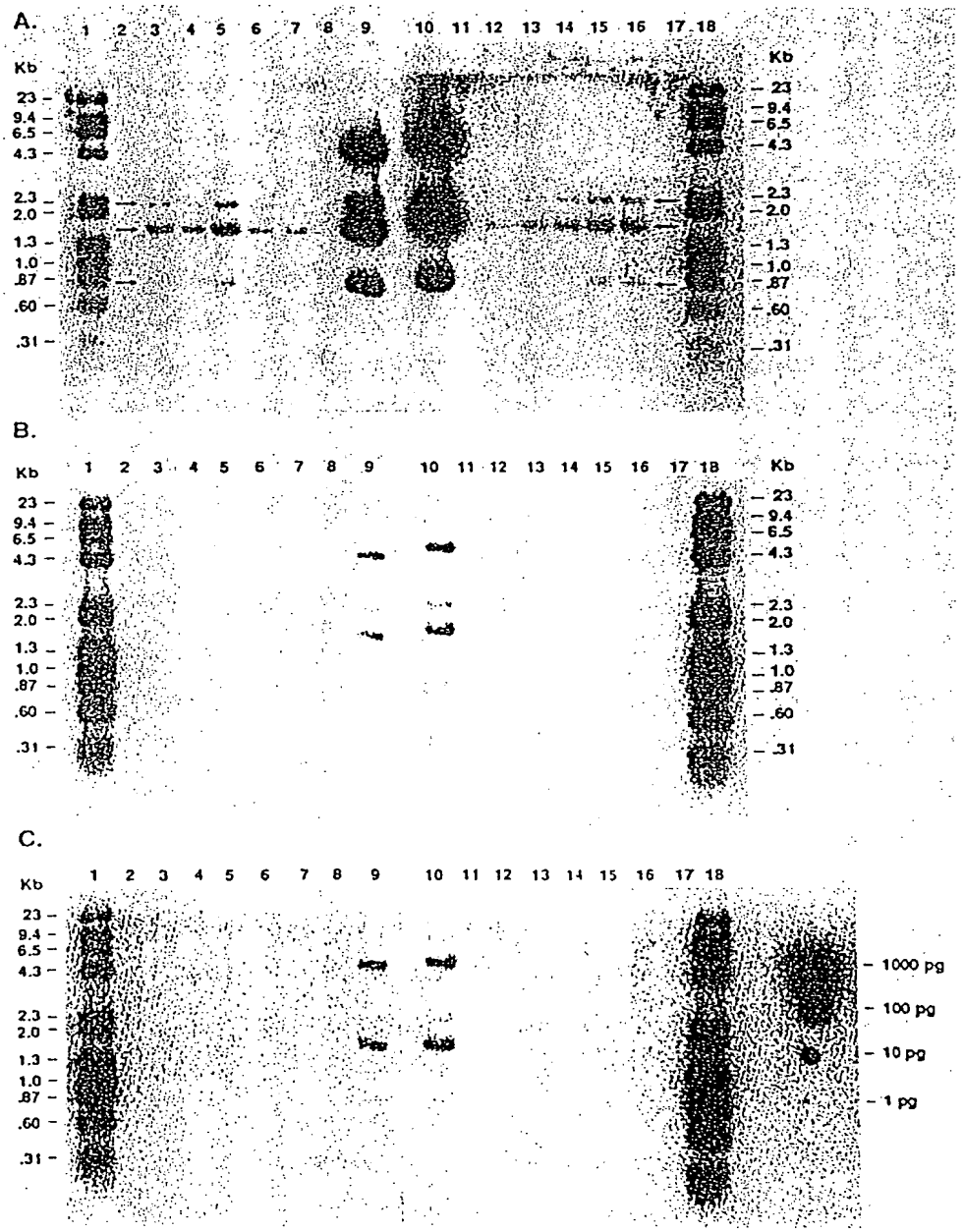


FIG. 4. Southern blot analysis of DNA from 7-day chicken embryos injected with 20 μl of SW272/cGH vector before incubation. High-molecular-size DNA (15 μg) was digested wtih *Bam*HI before analysis. The same filter was hybridized to three different probes: pSW272 probe (A), probe removed from panel A (B), virus-specific probe (C), probe removed from panel C (D), and cGH-specific probe (E). Probe hybridized to vector DNA in lanes 9 and 10 of panel A could not be completely removed. Sequences recognized by these probes are illustrated in Fig. 1. Lanes: 1 and 18, *Hin*dIII-digested lambda phage DNA, *Hae*III-digested ϕX174 DNA, and *Bam*HI-digested uninjected chicken blood DNA; 2 and 11, DNA from uninjected embryos; 8 and 17, DNA from blood of uninjected chickens; 3 to 7 and 12 to 16, DNA from vector-injected embryos; 9 and 10, *Bam*HI-digested DNA of pSW272/cGH (1 ng) plus uninjected chicken blood DNA. *Bam*HI fragments internal to the proviral vector are marked with arrows in panel A. *Bam*HI fragments containing the endogenous cGH sequence are marked by asterisks in panel E. Dot blot on the right of panel C contains the indicated amounts of pSW253 containing the REV-A provirus (5). Sizes are shown in kilobase pairs (Kb).

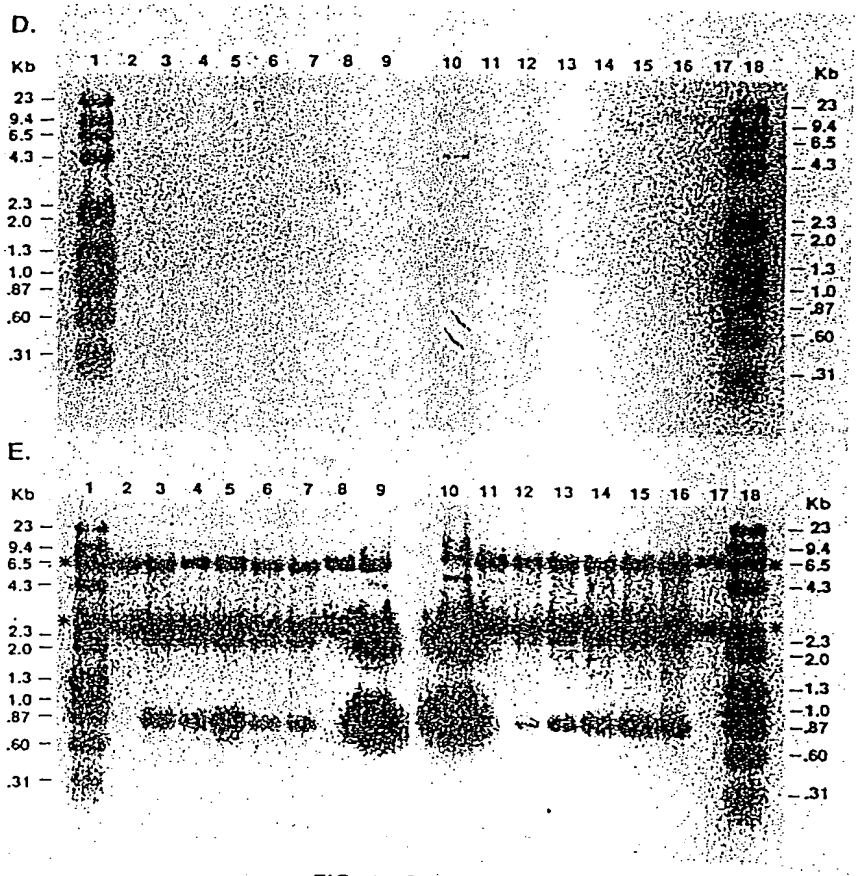


FIG. 4—*Continued.*

with radiolabeled DNA probes for the cGH coding sequence, 5′ and 3′ vector-specific sequences, or REV virus probe (see Fig. 1). Lane 1 of each panel contains a mixture of bacteriophage lambda and ϕX174 DNAs digested with *Hin*dIII and *Hae*III, respectively, and *Bam*HI-digested control chicken blood DNA. Lane 9 in Fig. 6A to C contains *Bam*HI-digested control chicken blood DNA and plasmid DNA of pSW272/cGH. Internal *Bam*HI fragments of vector DNA are indicated by arrows. *Bam*HI fragments derived from the endogenous cGH gene are shown by the asterisks. Internal fragments derived from the complete REV provirus are marked by chevrons. Results obtained by hybridization with the cGH-specific probe are shown in Fig. 6A. *Bam*HI-digested DNAs from control semen and blood (lanes 3 and 6, respectively) contain endogenous fragments of ~2.7 and ~6.4 kb. In lanes 4 and 7, DNAs from the semen and blood, respectively, of the vector-positive male contain an additional 0.86-kb fragment which derives from the vector SW272/cGH and hybridizes to the cGH probe. Although visible in the original autoradiogram, the 2.3-kb *Bam*HI fragment derived from the vector is not well resolved from the strongly hybridizing 2.7-kb *Bam*HI fragment derived from the endogenous cGH gene. Blood DNA appears to contain much less of the 0.86-kb fragment than does semen DNA.

Panel B of Fig. 6 shows results obtained when a similar filter was hybridized with the vector probe. Lanes 4 and 7 show that semen and blood DNA from an infected male contain *Bam*HI fragments of 0.86 and 1.6 kb. These fragments derive from the 5′ and 3′ ends of the integrated vector DNA, respectively. The additional 2.3-kb internal *Bam*HI fragment of vector DNA containing HSV-1 *tk* sequences does not hybridize to the vector probe used in Fig. 6B nor does *Bam*HI-digested DNA from semen and blood of uninfected control birds (lanes 3 and 6). No 1.4-kb fragment characteristic of replicating REV was observed. The patterns of semen and blood *Bam*HI DNA fragments hybridizing with these probes are similar to each other and are consistent with the pattern observed in *Bam*HI-digested DNA from infected embryos (Fig. 4).

Panel C of Fig. 6 shows results of hybridization with a virus-specific probe. Lanes 1 to 9 are as described for panels A and B. Lane 10 is blank. Lanes 11 and 12 contain *Bam*HI-digested plasmids pSW279 (39) and pSW253 (5), respectively. DNA in lane 11 has a 5′ LTR derived from SNV (with a *Bam*HI site) but a 3′ LTR derived from REV-A (without a *Bam*HI site). This provirus also lacks a 310-base-pair packaging sequence (E) located near the 5′ end of the provirus. The expected fragments generated by *Bam*HI digestion of this DNA are present at ~1.1 (E^-), ~1.8, ~2.2, and ~0.7 kb (visible at longer exposure times). Plasmid pSW253 in lane 12 contains the REV-A provirus and lacks the *Bam*HI site present in the SNV LTR. *Bam*HI digestion of this DNA generates the observed ~1.8- and ~2.2-kb fragments. The large fragment of ~9 kb in lane 12 contains 5′ and 3′ portions of the provirus and a portion of the *gag* gene

TABLE 1. cGH levels in chicken embryo serum[a]

C3-44-injected embryos		Uninjected embryos	
Bird no.	Amt (ng/ml) of cGH	Bird no.	Amt (ng/ml) of cGH
1	51	31	<0.80
2	180	32	<0.80
3	100	33	<0.80
4	0.9	34	<0.80
5	41	35	<0.80
6	200	36	0.85
7	2.6	37	<0.80
8	80	38	<0.80
9	44	39	<0.80
10	106	40	<0.80
11	4.5	41	<0.80
12	18	42	1.2
13	8.6	43	1.0
14	1.1	44	<0.8
15	0.92	45	<0.8
16	2.2	46	<0.8
17	0.8	47	<0.8
18	254	48	<0.8
19	10.8	49	1.1
20	240	50	1.2
21	168	51	1.2
22	32	52	0.9
23	56	53	1.2
24	12	54	0.9
25	42	55	<0.8
26	0.86	56	<0.8
27	0.70	57	<1.1
28	3.4	58	<0.8
29	1.4	59	<0.8
30	0.76	60	<0.8
		61	<0.8
		62	<0.8
		63	<0.8
		64	<0.8
		65	<0.8

[a] Embryos of unincubated eggs were injected with 10 μl of medium from cultures of clone C3-44. After 15 days of incubation, serum from each embryo was assayed by RIA for cGH.

sequence. The 0.7-kb fragment is observed at longer exposure times. DNAs in lanes 3, 4, 6, 7, and 9 do not contain sequences detectable with probe derived from the structural genes of REV.

DISCUSSION

Early chicken embryo development. Fertilization and the first 24 h of chicken embryonic development occur in the oviduct and uterus, concomitant with the accretion of albumen and deposition of the eggshell. During this period, attempts at gene transfer into the embryo must allow for surgical removal after fertilization and either reintroduction to the oviduct or extensive artificial culture (25, 26). Both of these approaches are technically difficult. Alternatively, infection of the embryo just after ovaposition represents a strategy well suited to vector-mediated gene transfer. The embryo at this stage is composed of at least 10,000 cells arranged in a disk-shaped blastoderm, one to two cells thick and 2 to 3 mm in diameter (10, 20). The day-old blastoderm floats on the yolk above a fluid-filled subgerminal cavity.

Previous studies have provided a detailed description of early chicken embryo development (10, 20) and insights regarding the developmental potential of cells comprising the embryonic blastoderm of a freshly laid egg (9, 11, 12, 22). Separated posterior and anterior portions of very young unincubated blastoderms appear totipotent, with similar ability to form embryos in vitro (10). The slightly older blastoderm exhibits cells of both upper epiblastic and lower hypoblastic layers. Separated from the lower layer of cells, the upper epiblastic layer retains its pluripotency, regenerates a new hypoblastic layer, and can subsequently form an early-stage embryo in vitro. When dissociated and grown in culture, epiblast cells form structures resembling embryoid bodies formed by murine teratocarcinomas (22). The hypoblastic layer, in contrast, survives but does not form embryonic structures (22).

All of the above observations suggest that successful infection of the early blastoderm with REV vectors might result in gene transfer into pluripotent embryonic stem cells. However, the REVs are primarily exogenous viruses in the chicken (41). Even though infected dams can transmit the virus vertically to their offspring by shedding virus into the egg (42), nucleic acid sequences closely related to REV are not endogenous to the chicken genome (Fig. 4C). The biology of virus-host interactions may preclude stable insertion of REV sequences into the chicken genome under natural conditions. Insertion of complete REV proviruses into the chicken genome could adversely affect viability, but even defective proviruses appear to be absent from the chickens analyzed in this study.

Infection of unincubated chicken embryo blastoderms. We have used replication-defective REV vectors ME111 and SW272/cGH to test the feasibility of retrovirus-mediated gene transfer in the chicken. The C3 helper cell line has been used to generate titers of about 10^4 infectious units per ml. The ME111 vector carries the Tn*5* neomycin phosphotransferase gene and the HSV-1 *tk* gene and has been described previously (8). The vector SW272/cGH carries a cDNA sequence encoding the cGH mRNA and the HSV-1 *tk* gene (see Fig. 1). Clone C3-44 released about 10^4 TKTU/ml and expressed about 500 ng of cGH per ml of culture medium. Analysis by RIA (35) (data not shown) and Western blotting (3) (Fig. 2) showed that cGH released by C3-44 and transduced by SW272/cGH is similar to natural cGH.

Glass needles (40 to 60 μm diameter) were used to deposit medium containing vector directly above and below the surface of the unincubated embryonic blastoderm. This method resulted in successful transduction of vector sequences into recipient embryos. Estimates of the amount of vector injected into the space beneath the blastoderm are based on the titer on BRLtk$^-$ cells ($\sim 10^4$ TKTU/ml) and the observation that REV titer on chicken cells could be 10- to 100-fold higher (39). We estimate that between 10^3 and 10^4 TKTU were injected per embryo.

Vector DNA present in 7-day embryos. Dot blot analysis of 7-day embryo DNA shown in Fig. 3 indicated that about 50% of injected embryos contained detectable vector sequences. Three different radiolabeled probes were used in Southern blot analysis of high-molecular-mass DNA from 10 embryos to distinguish the REV structural genes, the SW272/cGH vector, and endogenous cGH sequences from each other (see Fig. 1). Since most infected embryonic cells are likely to have a single copy of vector, these blots indicate that a significant percentage of the embryonic cells may carry vector sequences 7 days after infection. This is most evident in comparisons of endogenous and vector-specific *Bam*HI fragments hybridizing to the cGH probe (Fig. 4E). The lack of *Bam*HI fragments specific for replicating REV (Fig. 4C) confirms that gene transfer is primarily the result of the replication-defective REV vector and not of contaminating helper virus (17). These results show that early embryonic

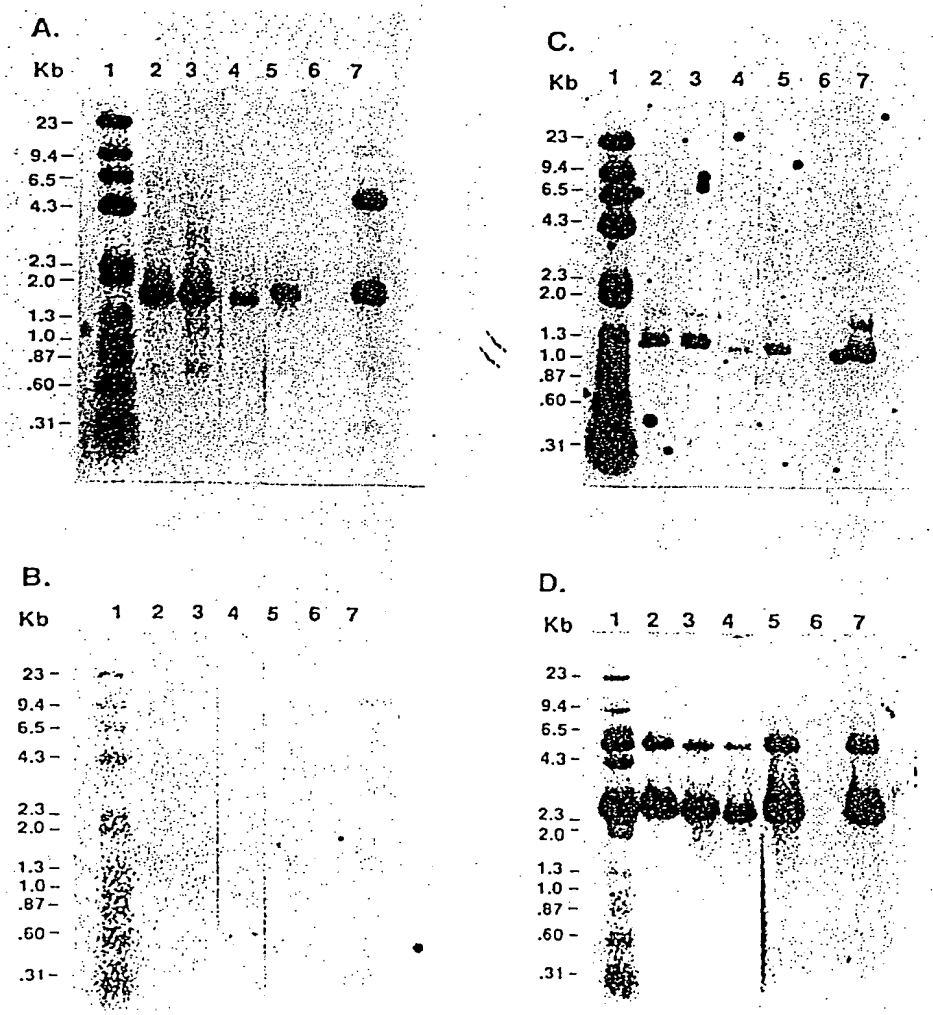


FIG. 5. Southern blot analysis of ~20 μg of *Bam*HI- and *Bgl*II-digested DNA from tissues of ME111-positive adult male 87725. (A through D) Replicate blots hybridized with radiolabeled vector probe (A), virus probe (B), *tk* probe (C), and cGH probe (D). Lanes: 1, *Hin*dIII-digested lambda phage DNA, *Hae*III-digested ϕX174 DNA and *Bam*HI- and *Bgl*II-digested negative control chicken blood DNA; 2, blood DNA; 3, brain DNA; 4, muscle DNA; 5, testis DNA; 6, blank; 7, *Bam*HI- and *Bgl*II-digested pME111 (50 pg) and negative control chicken blood DNA. Sizes (in kilobases) are shown at the left of each panel.

cells are susceptible to REV infection and that they persist during development, comprising a significant fraction of the 7-day embryo.

Expression of cGH in vector-injected embryos. Expression of the endogenous cGH gene occurs late during embryonic development. Caudal cells of the pituitary do not contain immunodetectable cGH until day 12 of embryonic development (19), whereas detectable plasma cGH does not appear until day 17 of incubation (15). Furthermore, response of the pituitary to the cGH secretagogue, thyrotrophin-releasing hormone, is not seen until hatching (7). The absence of endogenous REV and the restricted location and timing of endogenous cGH expression facilitate the distinctions between endogenous and vector-encoded genes and their products.

Expression of the cGH gene in vivo resulted in elevated serum cGH levels in about 50% of injected embryos when measured after 15 days of development (Table 1). Levels of serum cGH in 30 injected embryos varied from <1 ng/ml to 254 ng/ml, whereas none of the 35 uninjected controls had serum cGH levels above 2 ng/ml. Immunoprecipitated serum cGH from infected embryos comigrated with purified recombinant cGH as shown by Western blot analysis (Fig. 2). The relative contribution of somatic tissues to circulating levels of cGH is not known. These results are consistent with infection of embryonic stem cells present in the blastoderm at the time of vector injection and expression of vector-encoded cGH.

Vector DNA in tissues of adult males. Southern blot analysis of DNA from an adult male injected as an embryo with ME111 demonstrated the presence of vector in blood, brain, muscle, and testes (Fig. 5). Analysis of semen DNA by Southern blotting confirmed the presence of integrated unrearranged vector sequences in a low percentage of the sperm cells from a bird injected with SW272/cGH (Fig. 6). The pattern of *Bam*HI restriction fragments observed (0.86 and 1.6 kb) is consistent with that seen in Southern blot analysis of DNA from infected embryos. Probe containing HSV-1 *tk* sequences revealed the additional 2.3-kb *Bam*HI

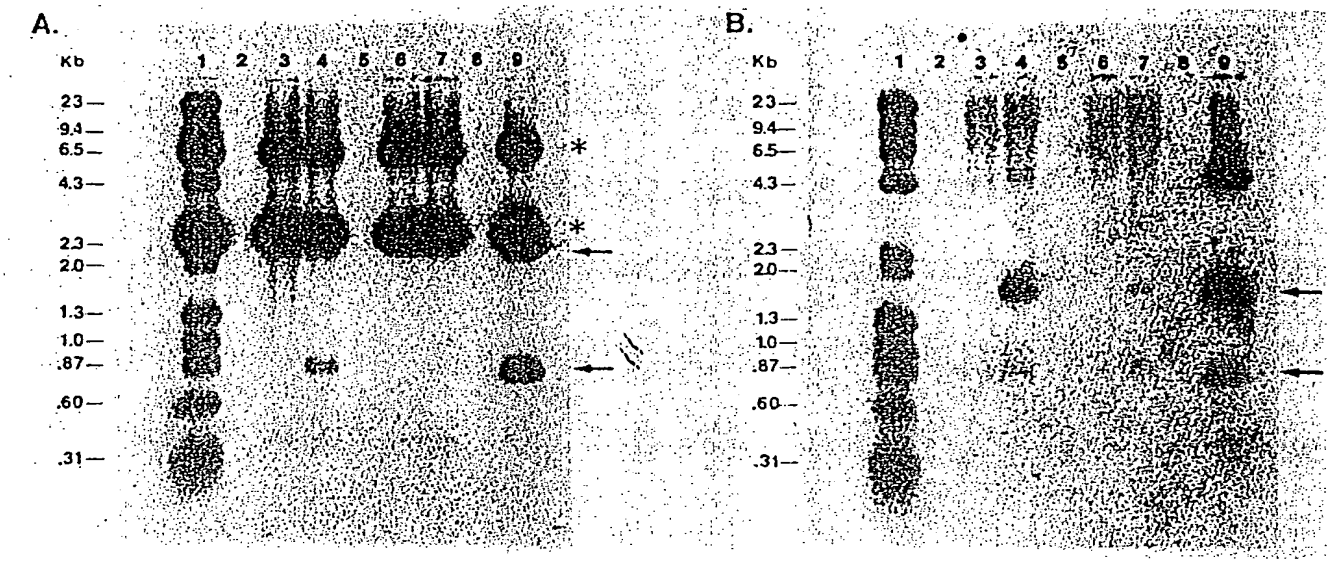


FIG. 6. Southern blot analysis of *Bam*HI-digested DNA from semen and blood of SW272/cGH-infected and control chickens. (A and B) Duplicate blots hybridized with radiolabeled probe specific for cGH or 5′ and 3′ regions of the vector, respectively. See the legend to Fig. 1 for a detailed description of probes. Lanes: 1, *Hin*dIII-digested lambda phage DNA, *Hae*III-digested ϕX174 DNA, and *Bam*HI-digested control chicken blood DNA; 3, control semen DNA; 4, SW272/cGH-infected G_0 semen DNA; 6, control blood DNA; 7, SW272/cGH-infected G_0 blood DNA; 9, *Bam*HI-digested plasmid pSW272/cGH (20 pg) and control chicken blood DNA. Arrows, internal *Bam*HI fragments of vector DNA; asterisks, endogenous *Bam*HI fragments derived from the cGH gene. (C) Hybridization with virus-specific probe. Lanes: 1 to 9, same as in panels A and B; 10, blank; 11 and 12, *Bam*HI-digested plasmids pSW279 and pSW253, which carry a hybrid SNV/REV-A provirus and REV-A provirus, respectively.

vector DNA fragment (data not shown) also seen in Fig. 5A. Vector sequences present in semen are not caused by contaminating blood cells since blood contains lower levels of vector DNA per microgram of total DNA, as shown by Southern blotting. Furthermore, blood cells were not detected in vector-positive semen subjected to microscopic examination nor could vector sequences be detected in negative control semen containing 1% vector-positive blood from a different bird. We did not observe any consistent *Bam*HI fragments representing junctions between cellular and vector sequences, probably because of the polyclonal makeup and low overall percentage of cells carrying the vector. No 1.4-kb *Bam*HI DNA fragment characteristic of replicating SNV was observed (40) (see Fig. 1).

Previous work with replication-competent derivatives of Rous sarcoma virus showed that infection of unincubated chicken embryos resulted in germ line insertion of proviral DNA (28, 29). In contrast, the same approach using competent REV resulted in somatic infection but did not lead to germ line insertion of proviral DNA (29). Similarly, follicular injection of REV vectors resulted in the presence of vector sequences in somatic cells, but germ line transmission of these sequences was not demonstrated (32). We recently confirmed germ line transmission of the replication-defective REV vector ME111 administered as described here to chicken embryos (1). Breeding studies are now in progress to determine whether semen from the chickens infected as embryos with defective REV vectors encoding cGH can accomplish germ line transmission of the vector DNA to progeny.

Conclusion. Replication-defective REV vectors can introduce new genetic information into the chicken by infecting somatic stem cells of the embryo. Susceptibility of these stem cells to infection by REV vectors provides another approach to the in vivo study of avian development (4) and vector-mediated gene expression. The possible applications of this technology are numerous.

ACKNOWLEDGMENTS

The authors thank Howard Temin for kindly providing the C3 helper cells and the SW272 and ME111 vectors and Norman Davidson for his continuing interest in this work. We also thank Joan Bennett and Julie Heuston for secretarial and drafting assistance.

LITERATURE CITED

1. Bosselman, R. A., R.-Y. Hsu, T. Boggs, S. Hu, J. Bruszewski, S. Ou, L. Kozar, F. Martin, C. Green, F. Jacobson, M. Nicolson, J.

Schultz, K. Semon, W. Rishell, and R. G. Stewart. 1989. Germline transmission of exogenous genes in the chicken. Science 243:533–535.
2. Bosselman, R. A., R.-Y. Hsu, J. Bruszewski, S. Hu, F. Martin, and M. Nicolson. 1987. Replication-defective chimeric helper proviruses and factors affecting generation of competent virus: expression of Moloney murine leukemia virus structural genes via the metallothionein promoter. Mol. Cell. Biol. 7:1797–1806.
3. Burnette, W. N. 1981. Western blotting: electrophoretic transfer of proteins from sodium dodecyl sulphate-polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radioiodinated protein A. Anal. Biochem. 112: 195–203.
4. Cepko, C. 1988. Retrovirus vectors and their applications in neurobiology. Neuron 1:345–353.
5. Chen, I. S. Y., T. W. Mak, J. J. O'Rear, and H. M. Temin. 1981. Characterization of reticuloendotheliosis virus strain T DNA and isolation of a novel variant of reticuloendotheliosis virus strain T by molecular cloning. J. Virol. 40:800–811.
6. Cone, R. D., and R. Mulligan. 1984. High efficiency gene transfer into mammalian cells: generation of helper-free recombinant retrovirus with broad mammalian host range. Proc. Natl. Acad. Sci. USA 81:6349–6353.
7. Decuypere, E., and C. G. Scanes. 1983. Variation in the release of thyroxine, triiodothyronine and growth hormone in response to thyrotrophin releasing hormone during development of the domestic fowl. Acta Endocrin. 102:220–223.
8. Emerman, M., and H. M. Temin. 1984. Genes with promoters in retrovirus vectors can be independently suppressed by an epigenetic mechanism. Cell 39:459–467.
9. Eyal-Gilaldi, H. 1984. The gradual establishment of cell commitments during the early stages of chick development. Cell Differ. 14:245–255.
10. Eyal-Giladi, H., and S. Kochav. 1976. From cleavage to primitive streak formation: a complementary normal table and a new look at the first stages of the development of the chick. Dev. Biol. 49:321–337.
11. Eyal-Giladi, H., S. Kochav, and M. K. Menashi. 1976. On the origin of primordial germ cells in the chick embryo. Differentiation 6:13–16.
12. Eyal-Giladi, H., and N. T. Spratt, Jr. 1965. The embryo-forming potencies of the young chick blastoderm. J. Embryol. Exp. Morph. 13:267–273.
13. Feinberg, A. P., and B. Vogelstein. 1983. A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal. Biochem. 132:6–13.
14. Graham, F. L., and A. J. Van der Eb. 1973. A new technique for the assay of infectivity of human adenovirus in DNA. J. Virol. 52:456–467.
15. Harvey, S., T. F. Davison, and A. Chadwick. 1979. Ontogeny of growth hormone and prolactin secretion in the domestic fowl. Gen. Comp. Endocrinol. 39:270–273.
16. Hayward, W. S., B. G. Neel, and S. M. Astrin. 1981. ALV-induced lymphoid leukosis: activation of a cellular *onc* gene by promoter insertion. Nature (London) 290:475–480.
17. Hu, S., J. Bruszewski, M. Nicolson, J. Tseng, R.-Y. Hsu, and R. Bosselman. 1987. Generation of competent virus in the REV helper cell line C3. Virology 159:446–449.
18. Hughes, S. H., E. Kosik, A. M. Fadly, D. W. Salter, and L. B. Crittenden. 1986. Design of retroviral vectors for the insertion of new information into the avian germ line. Poult. Sci. 65: 1459–1467.
19. Jozsa, R., C. G. Scanes, S. Vigh, and B. Mess. 1979. Functional differentiation of the embryonic chicken pituitary gland studied by immunohistological approach. Gen. Comp. Endocrinol. 39: 158–163.
20. Kochav, S., M. Ginsburg, and H. Eyal-Giladi. 1980. From cleavage to primitive streak formation: a complementary normal table and a new look at the first stages of the development of the chick. Dev. Biol. 79:296–308.
21. Miller, A. D., and C. Buttimore. 1986. Redesign of retrovirus packaging cell lines to avoid recombination leading to helper virus production. Mol. Cell. Biol. 6:2895–2902.
22. Mitrani, E., and H. Eyal-Giladi. 1982. Cells from early chick embryos in culture. Differentiation 21:56–61.
23. Moscovici, C., M. G. Moscovici, and H. Jimenez. 1977. Continuous tissue culture cell lines derived from chemically induced tumors of Japanese quail. Cell 11:95–103.
24. Noori-Daloii, M. R., R. A. Swift, H.-J. Kung, L. B. Crittenden, and R. L. Winter. 1981. Specific integration of REV proviruses in avain bursal lymphomas. Nature (London) 294:574–576.
25. Perry, M. M. 1987. A complete culture system for the chick embryo. Nature (London) 331:70–72.
26. Rowlett, K., and K. Simkiss. 1987. Explanted embryo culture—*in vitro* and *in ovo* techniques for domestic fowl. Br. Poult. Sci. 28:91–101.
27. Salter, D. W. and L. B. Crittenden. 1987. Chickens transgenic for a defective recombinant avian leukosis proviral insert express subgroup A envelope glycoprotein. Poult. Sci. 66:170.
28. Salter, D. W., E. J. Smith, S. H. Hughes, S. E. Wright, and L. B. Crittenden. 1987. Transgenic chickens: insertion of retroviral genes into the chicken germ line. Virology 157:236–240.
29. Salter, D. W., E. J. Smith, S. H. Hughes, S. E. Wright, A. M. Fadly, R. L. Witter, and L. B. Crittenden. 1986. Gene insertion into the chicken germ line by retroviruses. Poult. Sci. 65: 1445–1458.
30. Scanes, C. G., and T. J. Lauterio. 1984. Growth hormone: its physiology and control. J. Exp. Zool. 232:443–452.
31. Sevoian, M., R. N. Larose, and D. M. Chamberlain. 1964. Avian lymphomatosis. VI. A virus of unusual potency and pathogenicity. Avian Dis. 8:336–347.
32. Shuman, R. M., and R. N. Shoffner. 1986. Gene transfer by avian retroviruses. Poult. Sci. 65:1437–1444.
33. Sorge, J., and S. H. Hughes. 1982. Splicing of intervening sequences introduced into an infectious retroviral vector. J. Mol. Appl. Genet. 1:547–559.
34. Southern, E. M. 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. 98:503–517.
35. Souza, L. M., T. C. Boone, D. Murdock, K. Langley, J. Wypych, D. Fenton, S. Johnson, P. H. Lai, R. Everett, R.-Y. Hsu, and R. Bosselman. 1984. Application of recombinant DNA technologies to studies on chicken growth hormone. J. Exp. Zool. 232: 465–473.
36. Stoker, A. W., and M. J. Bissell. 1988. Development of avian sarcoma and leukosis virus-based vector-packaging cell lines. J. Virol. 62:1008–1015.
37. Sugden, B., K. Marsh, and J. Yates. 1985. A vector that replicates as a plasmid and can be efficiently selected in B lymphocytes transformed by Epstein-Barr virus. Mol. Cell. Biol. 5:410–413.
38. Swift, R. A., C. Boerkoel, A. Ridgway, D. J. Fujita, J. B. Dodgson, and H.-J. Kung. 1987. B lymphoma induction by reticuloendotheliosis virus: characterization of a mutated chicken synctial virus provirus involved in c-*myc* activation. J. Virol. 61:2084–2090.
39. Watanabe, S., and H. M. Temin. 1982. Encapsidation sequences for spleen necrosis virus, an avian retrovirus, are between the 5' long terminal repeat and the start of the *gag* gene. Proc. Natl. Acad. Sci. USA 79:5986–5990.
40. Watanabe, S., and H. M. Temin. 1983. Construction of a helper cell line for avian reticulendotheliosis virus cloning vectors. Mol. Cell. Biol. 3:2241–2249.
41. Witter, R. L., and D. C. Johnson. 1985. Epidemiology of reticuloendotheliosis virus in broiler breeder flocks. Avian Dis. 29:1140–1154.
42. Witter, R. L., E. J. Smith, and L. B. Crittenden. 1980. Tolerance, viral shedding, and neoplasia in chickens infected with non-defective reticuloendotheliosis viruses. Avian Dis. 25: 374–394.

Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities

PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
LinkOut
My NCBI (Cubby)

Related Resources
Order Documents
NLM Catalog
NLM Gateway
TOXNET
Consumer Health
Clinical Alerts
ClinicalTrials.gov
PubMed Central

1: Biotechnology (N Y). 1991 Sep;9(9):835-8.

Related Articles, Links

Transgenic production of a variant of human tissue-type plasminogen activator in goat milk: generation of transgenic goats and analysis of expression.

Ebert KM, Selgrath JP, DiTullio P, Denman J, Smith TE, Memon MA, Schindler JE, Monastersky GM, Vitale JA, Gordon K.

Tufts University School of Veterinary Medicine, North Grafton, MA 01536-1895.

We report the first successful production of transgenic goats that express a heterologous protein in their milk. The production of a glycosylation variant of human tPA (LAtPA--longer acting tissue plasminogen activator) from an expression vector containing the murine whey acid promoter (WAP) operatively linked to the cDNA of a modified version of human tPA was examined in transgenic dairy goats. Two transgenic goats were identified from 29 animals born. The first animal, a female, was mated and allowed to carry the pregnancy to term. Milk was obtained upon parturition and was shown to contain enzymatically active LAtPA at a concentration of 3 micrograms/ml.

PMID: 1367544 [PubMed - indexed for MEDLINE]

1: Biotechnology (N Y). 1991 Sep;9(9):839-43.

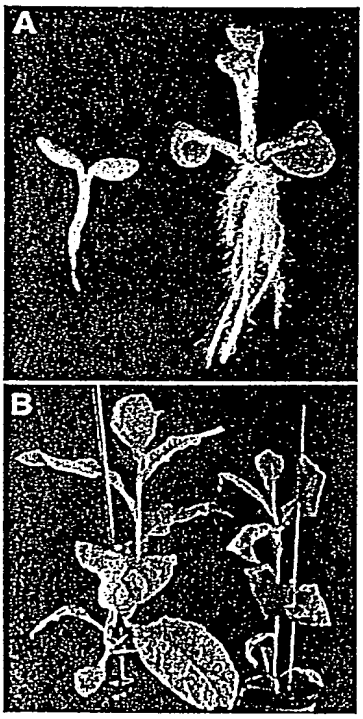
Transgenic expression of a variant of human tissue-type plasminogen activator in goat milk: purification and characterization of the recombinant enzyme.

Denman J, Hayes M, O'Day C, Edmunds T, Bartlett C, Hirani S, Ebert KM, Gordon K, McPherson JM.

Genzyme Corporation, Framingham, MA 01701.

A glycosylation variant of human tissue-type plasminogen activator (tPA) designated longer-acting tissue-type plasminogen activator (LAtPA) was extensively purified from the milk of a transgenic goat by a combination of acid fractionation, hydrophobic interaction chromatography and immunoaffinity chromatography. This scheme provided greater than 8,000-fold purification of the protein, a cumulative yield of 25% and purity greater than 98% as judged by SDS gel electrophoresis. SDS gel electrophoresis revealed that the transgenic enzyme was predominantly the "two chain" form of the protease. The specific activity of the purified transgenic protein, based on the average of the values obtained for three different preparations, was 610,000 U/mg as judged by amidolytic activity assay. This was approximately 84% of the value observed for the recombinant enzyme produced in mouse C127 cells. Analysis of the transgenic protein indicated that it had a significantly different carbohydrate composition from the recombinant enzyme produced in C127 cells. Molecular size analysis of the oligosaccharides from the transgenic and C127 cell-derived LAtPA preparations confirmed their differences and showed that the mouse cell-derived preparation contained larger, complex-type N-linked oligosaccharide structures than the material produced in goat mammary tissue.

PMID: 1367545 [PubMed - indexed for MEDLINE]



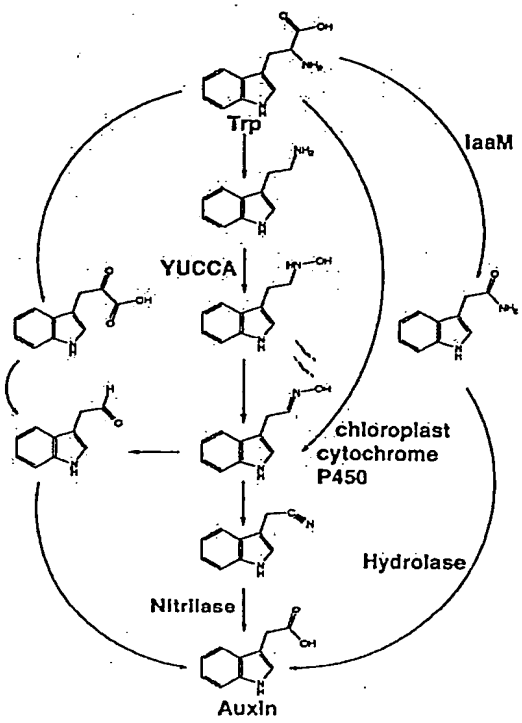


Fig. 4. (Left) YUCCA is involved in tryptophan-dependent auxin biosynthesis, and the YUCCA pathway is functional in other plants. (A) *yucca* is less sensitive to toxic tryptophan analogs. Wild-type (left) and *yucca* seedlings were grown on 0.5X MS medium containing 100-μM 5-mT for 10 days. (B) Comparison of wild-type (left) and transgenic tobacco plants overexpressing *YUCCA*. Fig. 5. (Right) YUCCA catalyzes a key step in auxin biosynthesis. Putative tryptophan-dependent auxin biosynthesis pathways and intermediates are shown (*2*). The indole-3-acetaldoxime intermediate was proposed recently (*25*).

may yield additional clues that can be used to elucidate the physiological roles of their mammalian counterparts.

References and Notes

1. P. J. Davies, Ed., in *Plant Hormones: Physiology, Biochemistry and Molecular Biology*, (Kluwer Academic, Dordrecht, Netherlands, 1995), pp. 1–12.
2. B. Bartel, *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 48, 51 (1997).
3. ———, G. R. Fink, *Proc. Natl. Acad. Sci. U.S.A.* 91, 6649 (1994).
4. D. Bartling, M. Seedorf, A. Mithofer, E. W. Weiler, *Eur. J. Biochem.* 205, 417 (1992).
5. J. Normanly, P. Grisafi, G. R. Fink, B. Bartel, *Plant Cell* 9, 1781 (1997).
6. R. C. Schmidt, A. Muller, R. Hain, D. Bartling, E. W. Weiler, *Plant J.* 9, 683 (1996).
7. J. J. King, D. P. Stimart, R. H. Fisher, A. B. Bleecker, *Plant Cell* 7, 2023 (1995).
8. M. Delarue, E. Prinsen, H. V. Onckelen, M. Caboche, C. Bellini, *Plant J.* 14, 603 (1998).
9. W. Boerjan *et al.*, *Plant Cell* 7, 1405 (1995).
10. A. Lehman, R. Black, J. R. Ecker, *Cell* 85, 183 (1996).
11. J. L. Celenza Jr., P. L. Grisafi, G. R. Fink, *Genes Dev.* 9, 2131 (1995).
12. D. Weigel *et al.*, *Plant Physiol.* 122, 1003 (2000).
13. C. P. Romano, P. R. Robson, H. Smith, M. Estelle, H. Klee, *Plant Mol. Biol.* 27, 1071 (1995).
14. R. J. Pitts, A. Cernac, M. Estelle, *Plant J.* 16, 553 (1998).
15. Analysis of free auxin was done as described by K. Chen, A. N. Miller, G. W. Patterson, and J. D. Cohen [*Plant Physiol.* 86, 822 (1988)]. Six-day-old *yucca* and wild-type seedlings grown in 0.5X MS medium at 22°C in white light were used for the analysis.
16. J. Mathur, C. Koncz, in *Arabidopsis Protocols*, J. M. Martinez-Zapater, J. Salinas, Eds. (Human Press, Totowa, NJ, 1998), vol. 82, pp. 31–34.
17. S. Sabatini *et al.*, *Cell* 99, 463 (1999).
18. C. P. Romano, M. B. Hein, H. J. Klee, *Genes Dev.* 5, 438 (1991).
19. D. M. Ziegler, *Drug Metab. Rev.* 19, 1 (1988).
20. J. R. Cashman, *Chem. Res. Toxicol.* 8, 165 (1995).
21. Y. Zhao, J. Chory, unpublished data.
22. Y. Zhao, S. Christensen, D. Weigel, J. Chory, unpublished data.
23. Y. Zhao, S. Christensen, J. Alonso, J. Ecker, J. Chory, unpublished data.
24. Y. Zhao, M. Neff, J. Chory, unpublished data.
25. A. K. Hull, R. Vij, J. L. Celenza, *Proc. Natl. Acad. Sci. U.S.A.* 97, 2379 (2000).
26. Transformation of tobacco was carried out as described by P. Gallois and P. Marinho [*Methods Mol. Biol.* 49, 39 (1995).]
27. N. B. Beaty, D. P. Ballou, *J. Biol. Chem.* 256, 4619 (1981).
28. ———, *J. Biol. Chem.* 256, 4611 (1981).
29. Recombinant YUCCA was purified from *Escherichia coli* as a maltose binding protein (MBP) fusion according to the procedures for expressing human FMOs [A. Brunelle *et al.*, *Drug Metab. Dispos.* 25, 1001 (1997)]. For activity assays, 350 μg of YUCCA-MBP was incubated with 2 mM tryptamine, 1 mM $NADP^+$, 1 mM glucose-6-phosphate, and 2.0 IU of glucose-6-phosphate dehydrogenase in 120-μL total volume at 37°C for 3 hours. The reactions were stopped by adding an equal volume of methanol. The control contained the same components, except methanol was added before the enzyme. The substrate and products were separated by thin-layer chromatography (TLC) with CH_2Cl_2/methanol/tetraethylammonium (TEA) (75:20:5). The product was eluted from the TLC plates, and Electrospray mass spectrometry was performed under positive mode to determine the molecular mass of the product.
30. Y. Zhao, J. R. Cashman, J. Chory, unpublished data.

31. We thank M. Estelle for providing *iaaL* overexpression *Arabidopsis* lines, L. Barden for assistance with the art work, T. Dabi for help with tobacco transformation, and J. Perry and members of the Chory lab for useful comments. Supported by grants from NIH (2R01GM52413) and NSF (MCB9631390) to J.C., from NSF (MCB 9723823) to D.W., from NIH (R01GM36426) to J.R.C., from DOE (DE-FG02-00ER15079) and from Minnesota Agricultural Experiment Station and the Bailey Endowment for Environmental Horticulture to J.D.C., and from the Howard Hughes Medical Institute. Y.Z. is a HHMI Fellow of the Life Sciences Research Foundation; S.C. was partially supported by an NSF fellowship; C.F. was a fellow of the Human Frontier Science Program and the Swiss NSF. J.C. is an Associate Investigator of the HHMI.

12 October 2000; accepted 1 December 2000

Transgenic Monkeys Produced by Retroviral Gene Transfer into Mature Oocytes

A. W. S. Chan, K. Y. Chong, C. Martinovich, C. Simerly, G. Schatten*

Transgenic rhesus monkeys carrying the green fluorescent protein (GFP) gene were produced by injecting pseudotyped replication-defective retroviral vector into the perivitelline space of 224 mature rhesus oocytes, later fertilized by intracytoplasmic sperm injection. Of the three males born from 20 embryo transfers, one was transgenic when accessible tissues were assayed for transgene DNA and messenger RNA. All tissues that were studied from a fraternal set of twins, miscarried at 73 days, carried the transgene, as confirmed by Southern analyses, and the GFP transgene reporter was detected by both direct and indirect fluorescence imaging.

Although transgenic mice have been invaluable in accelerating the advancement of biomedical sciences (*1–5*), many differences between humans and rodents have limited their usefulness (*6–9*). The major obstacle in producing transgenic nonhuman primates has been the low efficiency of conventional gene transfer protocols. By adapting a pseu-

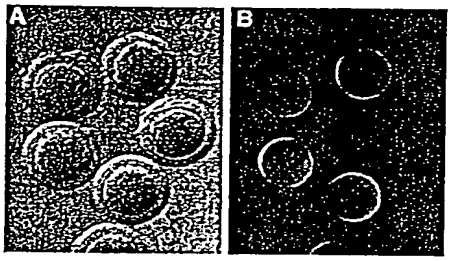


Fig. 1. Injection of VSV-G pseudotyped retroviral vector, enclosing the GFP gene and protein, into the perivitelline space of mature rhesus oocytes. (A) Transmitted light and (B) epifluorescence imaging of GFP carried within the vector particles. Magnification: ×100.

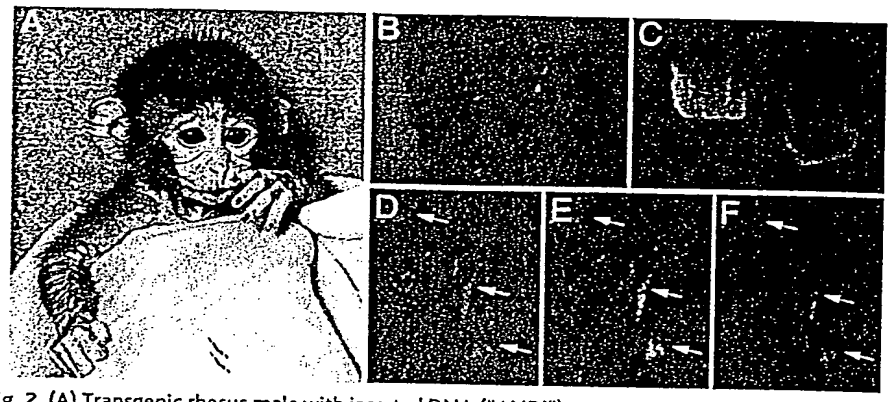


Fig. 2. (A) Transgenic rhesus male with inserted DNA ("ANDi"). GFP expression was observed in hair shafts (B) and toenails (C) by direct epifluorescent examination in the male stillborn but not in the accessible tissues from ANDi. Immunostaining and epifluorescent examination of placental frozen sections from the male stillborn demonstrates the presence of the GFP protein. (D) Anti-GFP detection in placenta by rhodamine (red) immunofluorescent microscopy. (E) GFP detection by fluorescein (green) epifluorescence of the same section demonstrates the direct expression of the transgene. (F) Overlay of the green (E) and red (D) images demonstrates colocalization of direct GFP fluorescence with anti-GFP imaging. Blue, Hoechst 33342 DNA staining. Magnification in (D) through (F): ×400.

dotyped vector system, efficient at up to 100% in cattle (*10, 11*), we circumvented problems in traditional gene transfer methodology to produce transgenic primates.

We injected 224 mature rhesus oocytes with high titer [10^8 to 10^9 colony-forming units (cfu)/ml] moloney retroviral vector pseudotyped with vesicular stomatitis virus envelope glycoprotein G (VSV-G pseudotype) into the perivitelline space (Fig. 1; Table 1; *10–12*). The VSV-G pseudotype carried the GFP gene under the control of either the cytomegalovirus early promoter (CMV) [referred to as LNCEGFP-(VSV-G)] or the human elongation factor-1 alpha promoter (hEF-1α) [referred to as LNEFEGFP-(VSV-G)] (*13*). Because ~10 to 100 pl was introduced into the perivitelline space, between 1 and 10 vector particles were introduced using LNCEGFP-(VSV-G) [10^9 cfu/ml] and between 0.1 to 1 with LNEFEGFP-(VSV-G) (10^8 cfu/ml). Oocytes were cultured for 6 hours before fertilization by intracytoplasmic sperm injection (ICSI). Vector particles incorporated into the oocyte in <4.5 hours as imaged by electron microscopy (*14*). Fifty-seven percent (n = 126) of embryos developed beyond the four-cell stage and 40 embryos were transferred to 20 surrogates, each carrying two embryos (Table 1). Rates for reproductive parameters are: fertilization [77% ICSI controls (*15*) versus 75% transgenesis], embryonic development [75% ICSI controls (*15*) versus 57% transgenesis], and implantation [66% ICSI controls (*16*) versus 25% transgenesis]. Most control ICSI pregnancies result in live offspring (83%) (*16*).

Five pregnancies resulted in the births of three healthy males (Table 1, Fig. 2). A set of fraternal twins miscarried at 73 days (150 to 155 days normal gestation) and a blighted pregnancy (implantation attempt without a fetus) also occurred. One fetal twin of the miscarriage was an anatomically normal male, while the other was largely resorbed in utero. The three births and the blighted pregnancy resulted from nine embryo transfers in which LNEFEGFP-(VSV-G) was used, whereas the twin pregnancy was established from 11 embryo transfers with LNCEGFP-(VSV-G) (Table 1).

Table 1. Transgenesis efficiency in rhesus embryos, fetuses, and offspring.

Construct	VSV-G pseudotype		Overall
	LNCEGFP	LNEFEGFP	
Eggs injected with vector	157	67	224
Eggs then injected with sperm	157	65	222
Fertilization rate	108 (69%)	58 (89%)	166 (75%)
Embryonic development of fertilized eggs	85 (79%)	41 (71%)	126 (76%)
Embryos transferred (two/surrogate)	22	18	40
Number of surrogates	11	9	20
Pregnancies/surrogate	1* (9%)	4 (44%)	5 (25%)
Fetal losses	2 (100%)	1 (25%)	3 (50%)
Births	0	3	3
Transgenic	2 of 2	1 of 4	3 of 6
Transgenic birth/embryos transferred	0	1 (5.5%)	1 (2.5%)
Transgenic birth/pregnancies	0	1 (25%)	1 (20%)

*Twin pregnancy.

Transgene integration, transcription, and expression from the newborns were examined in hair, blood, umbilical cords, placentae, cultured lymphocytes, buccal epithelial cells, and urogenital cells passed in urine, along with 13 tissues from the male stillborn, nine from the resorbed one, and specimens from the blighted pregnancy (*17*). Polymerase chain reaction (PCR) was performed with primer sets that covered the flanking region of the vector pLNC-EGFP or pLNEF-EGFP and the GFP gene (*18*). One newborn, ANDi, showed the presence of the transgene in all analyzed tissues, and the transgene was present in all tissues analyzed from both stillbirths including placentae and testes (Fig. 3). Total RNA was extracted for standard reverse transcription followed by PCR amplification (RT-PCR) with primer sets specific for the transgene (*18*). Transgene transcription was demonstrated in all of the tissues in the fetuses and in the accessible tissues from the infant carrying the transgene (Fig. 3).

Southern blot analysis of 10 tissues from the male stillbirth and eight samples from the other twin demonstrated multiple integration sites into their genomic DNA (Fig. 4) (*19*). Vector integration was determined by PCR of placenta, cord, blood, hair, and buccal cells using a primer set specific for the unique retroviral long terminal repeat (LTR) regions indicative of successful provirus integration into the host genome (*20, 21*). This provirus sequence was found in one infant and both

Oregon Regional Primate Research Center, Center for Women's Health, and Departments of Cell-Developmental Biology and Obstetrics-Gynecology, Oregon Health Sciences University, 505 NW 185th Avenue, Beaverton, OR 97006, USA.

*To whom correspondence should be addressed. E-mail: schatten@ohsu.edu

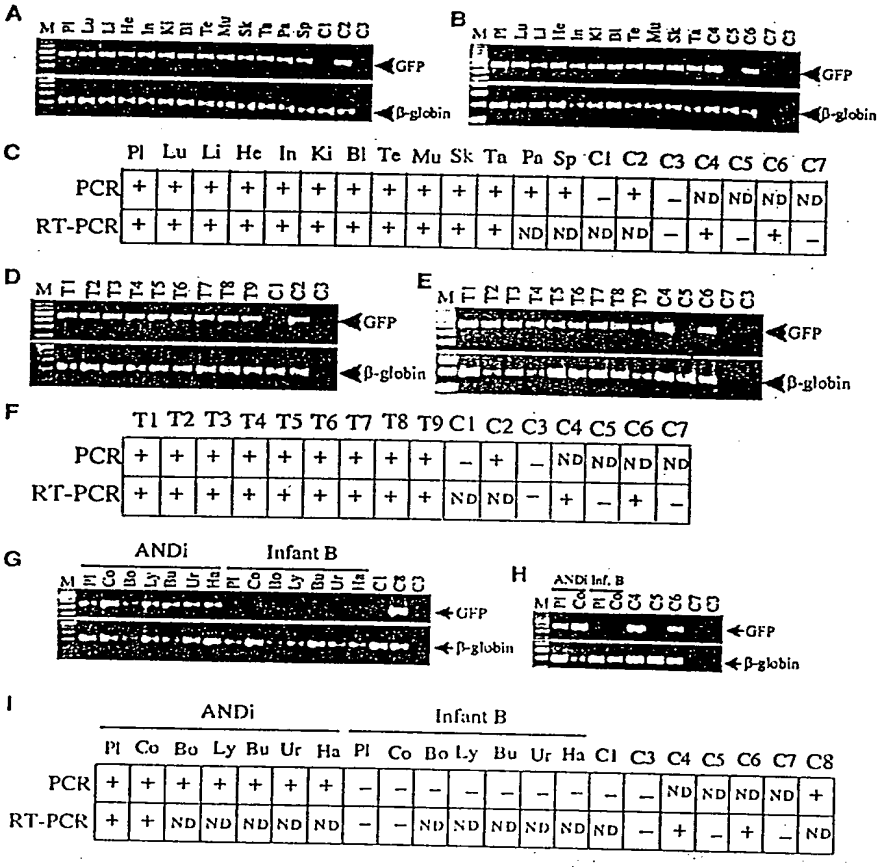


Fig. 3. PCR and RT-PCR analyses of transgenic and control tissues. (A) Thirteen tissues from an intact fetus were submitted for PCR and (B) 11 tissues for RT-PCR. (C) Analysis of the male stillborn. Tissues from the reabsorbed fetus were collected from eight different regions to ensure broad representation, because precise anatomical specification was limited. (D through F) PCR, RT-PCR of the reabsorbed fetus. A total of seven samples were obtained from each offspring for PCR (G), two samples for RT-PCR (H) from "ANDi" and one of the other two male offspring. (I) Analysis of the newborns, indicates that "ANDi" is a transgenic male with the presence of mRNA in all analyzed tissues. Co, cord; Bo, blood; Ly, lymphocyte; Bu, buccal cells; Ur, urine; Ha, hair; Pl, placenta; Lu, lung; Li, liver; He, heart; In, intestine; Ki, kidney; Bl, bladder; Te, testis; Mu, muscle; Sk, skin; Ta, tail; Pa, pancreas; Sp, spleen; T1 = placenta from reabsorbed fetus; T2 to T9 = tissues retrieved from eight regions of the reabsorbed fetus; C1 = nontransgenic rhesus tissue; C2 = C1 + pLNC-EGFP; C3 = ddH_2O; C4 = 293GP-LNCEGFP packaging cell; C5 = nontransgenic liver; C6 = transgenic lung without DNase; C7 = transgenic lung without reverse transcription; C8 = C1 + pLNEF-EGFP. ND, not determined.

stillbirths (Fig. 4D). Infant welfare considerations limited tissue availability, and genomic DNA obtained was insufficient for Southern analysis. The male infant with the inserted transgene has been named "ANDi" (for "inserted DNA," in a reverse transcribed direction; Fig. 2A).

GFP direct fluorescence in the toenails and hair of the fetus, as well as the placenta (Fig. 2, B through F), provided further evidence of transgenesis. Colocalization between direct GFP fluorescence and indirect anti-GFP immunocytochemical imaging demonstrated that the GFP protein is found exclusively at the direct fluorescence sources (Fig. 2, D through F). Furthermore, neither direct fluorescein nor indirect rhodamine fluorescence was observed in controls (*22*). Because tissues from the fetus originated from the three germ layers, the timing of transgene integration may have occurred before implantation, perhaps even before the first DNA replication cycle (*10*). The high efficiency of this approach has been linked to the absence of the nuclear envelope in oocytes naturally arrested in second meiotic metaphase (*10, 23*).

The miscarriage is likely due to the twin pregnancy, which is rare and high-risk in rhesus. The twin stillbirth originated from the

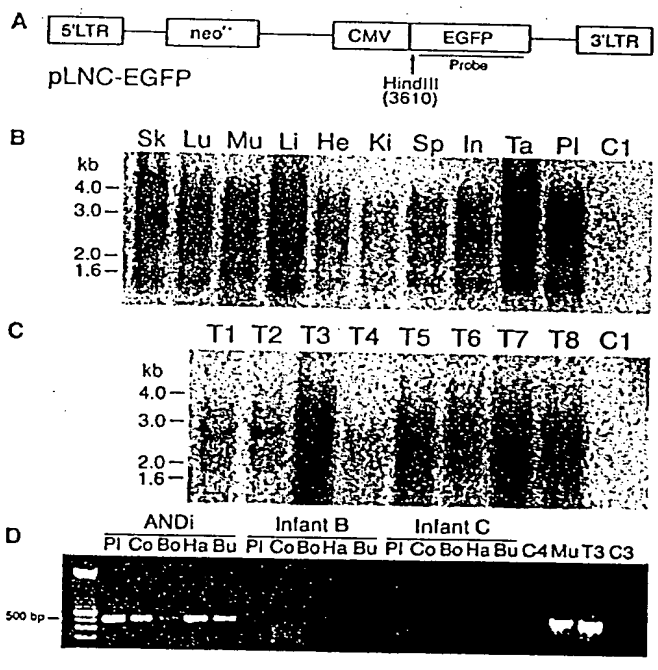


Fig. 4. (A) Southern blot analysis of Hind III (single digestion site) digested genomic DNA. Full-length GFP labeled with [^{32}P] was used as a probe to detect the transgene, which was detected in genomic DNA of the normal male stillbirth (B) and reabsorbed fetus (C). Nontransgenic rhesus tissue was used as a negative control (C1) and pLNC-EGFP DNA as a positive control (not shown). Various sized fragments were demonstrated in tissues obtained from each. This result indicates multiple integration sites due to the use of a restriction enzyme with a single digestion site within the transgene. (D) Detection of the unique provirus sequence. A total of five tissues from each infant and two tissues from a male stillbirth and the reabsorbed fetus were submitted for PCR. Provirus sequence was detected in "ANDi" and the two stillbirths (*42*), which indicates that they are transgenic. Abbreviations are the same as those in Fig. 3. Mu, muscle from the male stillborn; T3, tissue from the reabsorbed fetus.

higher titer vector, whereas the three births, including the transgenic one, and the blighted pregnancy originated from the lower titer LNEFEGFP-(VSV-G) vector (10^8 cfu/ml; Table 1). Although only one live offspring is shown to be transgenic, we cannot yet exclude the possibility of transgenic mosaics in the others. We have neither demonstrated germline transmission nor the presence of transgenic sperm; this must await ANDi's development through puberty in about 4 years. Vector titers and volume injected may play crucial roles in gene transfer efficiency. These offspring and their surrogates are now housed in dedicated facilities with ongoing, stringent monitoring.

Nonhuman primates are invaluable models for advancing gene therapy treatments for diseases such as Parkinson's (*24*) and diabetes (*25*), as well as ideal models for testing cell therapies (*26*) and vaccines, including those for HIV (*27, 28*). Although we have demonstrated transgene introduction in rhesus monkeys, significant hurdles remain for the successful homologous recombination essential for gene targeting (*29*). The molecular approaches for making clones [either by embryo splitting (*30*) or nuclear transfer (*31–36*)], utilizing stem cells (*37–39*), and now producing transgenic monkeys, could be combined to produce the ideal models to accelerate discoveries and to bridge the scientific gap between transgenic mice and humans.

References and Notes

1. M. J. Blouin *et al.*, *Nature Med.* 6, 177 (2000).
2. R. L. Eckert *et al.*, *Int. J. Oncol.* 16, 853 (2000).
3. H. M. Hsieh-Li *et al.*, *Nature Genet.* 24, 66 (2000).
4. A. M. Murphy *et al.*, *Science* 287, 488 (2000).
5. E. J. Weinstein *et al.*, *Mol. Med.* 6, 4 (2000).
6. J. A. Thomson, V. S. Marshall, *Curr. Topics Dev. Biol.* 38, 133 (1998).
7. A. W. S. Chan *et al.*, *Mol. Hum. Reprod.* 6, 26 (2000).
8. K. R. Chien, *J. Clin. Invest.* 98, S19 (1996).
9. R. P. Erickson, *BioEssays* 18, 993 (1996).
10. A. W. S. Chan *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 95, 14028 (1998).
11. J. K. Yee *et al.*, *Methods Cell Biol.* 43, 99 (1994).
12. The GFP vector was injected into the perivitelline space (*10*) of in vivo matured rhesus oocytes (*40*), fertilized by ICSI 6 hours later. Embryos at the four- to eight-cell stage were selected for embryo transfer on the basis of morphology. Surrogate females were selected on the basis of serum estradiol and progesterone levels (*15*).
13. The GFP gene from plasmid pEGFP-N1 was inserted into the retroviral vector pLNCX using standard recombinant DNA techniques [Web supplement 1 (*41*)].
14. Oocytes for electron microscopy were fixed in Ito-Karnovsky's fixative [Web supplement 2 (*41*)].
15. L. Hewitson *et al.*, *Hum. Reprod.* 13, 2786 (1998).
16. L. Hewitson *et al.*, *Nature Med.* 5, 431 (1999).
17. Genomic DNA was extracted from tissues obtained from the stillbirths [Web supplement 3 (*41*)].
18. PCR was performed using specific primers that amplify the flanking region of the GFP gene. Provirus was detected by using a primer set specific to the unique LTR region of genomic integrated virus. Transgene was detected by standard reverse transcription followed by PCR [Web supplement 4 (*41*)].
19. Southern analysis was performed using genomic DNA followed by restriction enzyme digestion using a unique site within the vector and detected by a GFP [^{32}P]-labeled probe [Web supplement 5 (*41*)].
20. B. Schott *et al.*, *Nucleic Acid Res.* 25, 2940 (1997).
21. N. Chinnasamy *et al.*, *Hum. Gene Ther.* 11, 1901 (2000).
22. Biopsied tissues were snap frozen, sectioned, fixed, and imaged with anti-GFP using rhodamine-conjugated anti-mouse (IgG) secondary antibody [Web supplement 6 (*41*)].
23. T. Roe *et al.*, *EMBO J.* 12, 2099 (1993).
24. J. H. Kordower *et al.*, *Science* 290, 767 (2000).
25. H. C. Lee *et al.*, *Nature* 408, 483 (2000).
26. D. H. Barouch *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 97, 4192 (2000).
27. M. J. Kuroda *et al.*, *J. Virol.* 74, 8751 (2000).
28. N. Nathanson *et al.*, *AIDS (suppl. A)* 13, S113 (1999).
29. U. Muller, *Mech. Dev.* 82, 1 (1999).
30. A. W. S. Chan *et al.*, *Science* 287, 317 (2000).
31. I. Wilmut *et al.*, *Nature* 385, 810 (1997).
32. J. B. Cibelli *et al.*, *Science* 280, 1256 (1998).
33. T. Wakayama *et al.*, *Nature* 394, 369 (1998).
34. D. P. Wolf *et al.*, *Biol. Reprod.* 60, 199 (1999).
35. A. Onishi *et al.*, *Science* 289, 1188 (2000).
36. I. A. Polejaeva *et al.*, *Nature* 407, 86 (2000).
37. J. A. Thomson *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 92, 7844 (1995).
38. M. J. Shamblott *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 95, 13726 (1998).
39. J. A. Thomson *et al.*, *Science* 282, 1145 (1998).
40. G. J. Wu *et al.*, *Biol. Reprod.* 55, 269 (1996).
41. Supplementary figures are available at www.sciencemag.org/cgi/content/full/291/5502/309/DC1.
42. A. W. S. Chan, K. Y. Chong, C. Martinovich, C. Simerly, G. Schatten, data not shown.

43. We thank J. C. Burns (University of California San Diego); B. True (University of Wisconsin–Madison); K. Wells (U.S. Department of Agriculture); Chiron Inc.; and all at the Oregon Regional Primate Research Center (ORPRC), especially M. Axthelm, J. Bassir, J. M. Cook, N. Duncan, M. Emme, J. Fanton, A. Hall, L. Hewitson, D. Jacob, E. Jacoby, A. Lewis, C. M. Luetjens, C. Machida, G. Macginnis, B. Mason, T. Swanson, D. Takahashi, K. Tice, J. Vidgoff, M. Webb, and S. Wong. Procedures approved by the Oregon Health Sciences University/ORPRC Animal Care and Biosafety Committees. Supported by NIH/National Center for Research Resources (NCRR) (ORPRC; M. S. Smith, Director) and grants (NCRR, National Institute of Child Health and Human Development to G.S.).

7 November 2000; accepted 14 December 2000

Categorical Representation of Visual Stimuli in the Primate Prefrontal Cortex

David J. Freedman,[1,2,5] Maximilian Riesenhuber,[3,4,5] Tomaso Poggio,[3,4,5] Earl K. Miller[1,2,5]*

The ability to group stimuli into meaningful categories is a fundamental cognitive process. To explore its neural basis, we trained monkeys to categorize computer-generated stimuli as "cats" and "dogs." A morphing system was used to systematically vary stimulus shape and precisely define the category boundary. Neural activity in the lateral prefrontal cortex reflected the category of visual stimuli, even when a monkey was retrained with the stimuli assigned to new categories.

Categorization refers to the ability to react similarly to stimuli when they are physically distinct, and to react differently to stimuli that may be physically similar (*1*). For example, we recognize an apple and a banana to be in the same category (food) even though they are dissimilar in appearance, and we consider an apple and a billiard ball to be in different categories even though they are similar in shape and sometimes color. Categorization is fundamental; our raw perceptions would be useless without our classification of items as furniture or food. Although a great deal is known about the neural analysis of visual features, little is known about the neural basis of the categorical information that gives them meaning.

In advanced animals, most categories are learned. Monkeys can learn to categorize stimuli as animal or non-animal (*2*), food or non-food (*3*), tree or non-tree, fish or non-fish (*4*), and by ordinal number (*5*). The neural correlate of such perceptual categories might be found in brain areas that process visual form. The inferior temporal (IT) and prefrontal (PF) cortices are likely candidates; their neurons are sensitive to form (*6–9*) and they are important for a wide range of visual behaviors (*10–12*).

The hallmark of perceptual categorization is a sharp "boundary" (*13*). That is, stimuli from different categories that are similar in appearance (e.g., apple/billiard ball) are treated as different, whereas distinct stimuli within the same category (e.g., apple/banana) are treated alike. Presumably, there are neurons that also represent such sharp distinctions. This is difficult to assess with a small subset of a large, amorphous category (e.g., food, human, etc). Because the category boundary is unknown, it is unclear whether neural activity reflects category membership or physical similarity.

[1]Center for Learning and Memory, [2]RIKEN-MIT Neuroscience Research Center, [3]Center for Biological and Computational Learning, [4]McGovern Institute for Brain Research, [5]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

*To whom correspondence should be addressed. E-mail: ekm@ai.mit.edu

Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities

PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
LinkOut
My NCBI (Cubby)

Related Resources
Order Documents
NLM Catalog
NLM Gateway
TOXNET
Consumer Health
Clinical Alerts
ClinicalTrials.gov
PubMed Central

1: Biotechnology (N Y). 1991 Sep;9(9):844-7. Related Articles, Links

Generation of transgenic dairy cattle using 'in vitro' embryo production.

Krimpenfort P, Rademakers A, Eyestone W, van der Schans A, van den Broek S, Kooiman P, Kootwijk E, Platenburg G, Pieper F, Strijker R.

Department of Embryology, Gene Pharming Europe B.V., Leiden, The Netherlands.

We have combined gene transfer, by microinjection, with 'in vitro' embryo production technology, enabling us to carry out non-surgical transfer, to recipient cows, of microinjected embryos that have been cultured from immature oocytes. Using this approach, we have established 21 pregnancies from which 19 calves were born. Southern blot analysis proved that in two cases the microinjected DNA had been integrated in the host genome.

PMID: 1367358 [PubMed - indexed for MEDLINE]

EXHIBIT MM

Gene transfer into sheep

Simons, J. Paul; Wilmut, Ian; Clark, A. John; Archibald, Alan L.; Bishop, John O.; Lathe, Richard

Bio/Technology (1988), 6(2), 179-83 CODEN: BTCHDA; ISSN: 0733-222X. English.

Six transgenic sheep were formed by microinjection of ONA into the pronuclei of single-cell eggs. Three DNA constructs were microinjected: (1) pMK, which contains the mouse metallothionein-1 (MT) promoter linked to the herpes simplex thymidine kinase gene nTK) (2) BLG-FIX, which contains the β-lactoglobulin gene (BLG) linked to cDNA sequences encoding for human blood-coagulation factor IX and (3) BLG-α1AT, which contains gene BLG linked to cDNA sequences encoding human α1-antitrypsin. The DNA in the transgenic sheep has not undergone rearrangement, as verified by hybridization assays. Hybridization intensities revealed the presence of single and multiple copies of constructs in the 6 lambs. Multiple copies had head-to-head and head-to-tail tandem arrangements. One of the offspring has the pMK construct, 4 of the offspring carry the BLG-FIX construct, and the last offspring carries the BLG-α1AT construct. Offspring from these transgenic sheep also carry the transgenic genes.

EXHIBIT NN

Expression of human anti-hemophilic factor IX in the milk of transgenic sheep

Clark, A. J.; Bessos, H.; Bishop, J. O.; Brown, P.; Harris, S.; Lathe, R.; McClenaghan, M.; Prowse, C.; Simons, J. P.; et al.

Bio/Technology (1989), 7(5), 487-92 CODEN: BTCHDA; ISSN: 0733-222X. English.

Transgenic livestock may prove useful for the large scale production of valuable proteins. By targeting expression to the mammary gland these proteins could be harvested from milk. To this end, a hybrid gene was designed to direct the synthesis of human anti-hemophilic factor IX to the mammary gland, and introduced into sheep. Two transgenic ewes, each carrying about 10 copies of the foreign gene, have been analyzed for expression. Both animals express human factor IX RNA in the mammary gland and secrete the corresponding protein into their milk.